Technische Universität Darmstadt

Analyse von Datei-Zugriffen zur Potentialermittlung für Information Lifecycle Management

Lars Arne Turczyk, Roswitha Gostner, Rainer Berbner,
Oliver Heckmann, Ralf Steinmetz

KOM Technical Report 01/2005
Version 1.0
Dezember 2005

Department of Electrical Engineering & Information Technology
Merckstraße 25 • D-64283 Darmstadt • Germany
Phone: +49 6151 166150
Fax: +49 6151 166152
Email: info@KOM.tu-darmstadt.de
URL: http://www.kom.tu-darmstadt.de/

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AO	Abgabenordnung
ATA	Advanced Technology Attachments
BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht
DMF	Daten-Management-Forum
GoB	Grundsätze der ordnungsgemäßen Buchführung SNIA Storage Networking Industry Association
GPDdU	Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen
HGB	Handelsgesetzbuch
ILM	Information Lifecycle Management
PACS	Picture Archive and Communications Systems
SAN	Storage Area Networks
SCSI	Small Computer System Interface
SLA	Service Level Agreement
TWG	Technical Work Groups
TC	Technical Council
TLG	Technical Liaison Group

1 Einleitung

Information Lifecycle Management (ILM) ist eine Strategie, die den Wert von Daten in Abhängigkeit eines Geschäftsmodells und dessen Dienstgütevereinbarungen (Service Level Agreements) klassifiziert [1] [2] [3]. Dadurch werden die Daten automatisch einem Speichermedium zugeordnet, so dass die vorhandenen Ressourcen bestmöglich ausgenutzt werden, was insgesamt zu einer Reduktion von Kosten führt.

Damit ILM angewandt werden kann, muss zuerst das Potential an migrierbaren Dateien identifiziert werden. Dies geschieht in dieser Arbeit exemplarisch anhand von Untersuchungen der Zugriffe auf Dateien über einen Zeitraum von 90 bis 180 Tagen.

Im Rahmen des Technical Reports wurde ein Datenbestand eines Unternehmens analysiert, um zu prüfen, ob genügend Potenzial vorhanden ist, so dass ILM nutzbringend angewendet werden kann.

In verschiedenen Publikationen wurden Dateizugriffe 1981, 1992 und 1998 analysiert [6] [7] [8]. Die vorliegende Analyse unterscheidet sich von den bekannten Arbeiten durch qualitative Abschätzungen eines Potentials für ILM. Dazu wird eine Unternehmensdatenbank untersucht, während bei den bekannten Studien Daten von Behörden, Universitäten und des Militärs betrachtet werden.

Die Ergebnisse der vorliegenden Analyse lassen sich als exemplarische Situation in eine Simulationsumgebung überführen. Dadurch liefert diese Arbeit die Möglichkeit, ILM-Szenarien mit realen Daten im Labor zu simulieren und weitere Erkenntnisse zu generieren.

Der vorliegende Bericht beginnt mit einer Einführung in ILM. In Kapitel 3 wird die Ausgangssituation der Untersuchung dargestellt. In Kapitel 4 wird die Datenbasis aus insgesamt neun Projekten identifiziert. In den anschließenden Kapiteln werden die Zugriffsdaten je Projekt und gesamt auf 90-Tage-Basis und auf 400-Tage-Basis ausgewertet. Die Kapitel 8 und 9 stellen das Potential für ILM dar und fassen die Ergebnisse der Studie zusammen.

2 Definition ILM

Das von der Speicherindustrie erarbeitete Konzept ILM soll die in Zukunft aufkommende Verwaltung von sehr großen Datenmengen handhaben können. Eine erste allgemein anerkannte Definition des Begriffes wurde im Jahr 2004 von der SNIA vorgestellt [2].

2.1 SNIA (Storage Networking Industry Association)

Mitte 1999 startete eine Gruppe führender Hersteller von Speicherlösungen eine Initiative zur Entwicklung von Standards für Multivendor Storage Area Networks (SANs), die SNIA [1, 2, 3]. Das gesetzte Ziel der SNIA ist es, Schlüsselfunktionalitäten zu konzipieren, Empfehlungen für ein verbessertes Speichermanagement zu erarbeiten und Mechanismen für die gemeinsame Nutzung von Speichertechniken und Daten zu entwickeln. Die von namhaften Herstellerfirmen erstellten Spezifikationen sollen die Kommunikation und das Management heterogener Speichernetze (SAN) verbessern.

2.2 Definition von ILM

Im Sommer 2004 stellte die SNIA ihre erarbeitete Definition des Begriffes ILM vor, welche in zwei Teile aufgeteilt wurde, um die gesamte Reichweite des Konzeptes zu erläutern:

> *ILM beinhaltet Policies, Prozesse, Erfahrungen und Werkzeuge, die den Wert einer Information innerhalb eines Geschäftsmodells mit der adäquaten und kosteneffizientesten IT-Infrastruktur verbinden und zwar von der Erstellung einer Information bis zur endgültigen Archivierung. Informationen sind abhängig von Geschäftsmodellen, Policy Management und Dienstgütevereinbarungen, welche mit Applikationen, Metadaten, [anderen] Informationen und Daten verknüpft sind.*
>
> *Quelle: [1], Eigene Übersetzung*

Mit „Policies, Prozessen, Erfahrungen und Werkzeugen“ wird die Vorgehensweise des Datenzentrums adressiert, in dem sich die klassische Übertragung und Archivierung von Daten zwischen verschiedenen hierarchischen Speichersystemen abspielt. Ebenfalls werden dort beschreibende Daten über Daten, so genannte Metadaten, erstellt.

Informationen werden innerhalb von ILM von der „Erstellung [...] bis zur endgültigen Archivierung“ betrachtet, das heißt, es wird für eine angemessene Infrastruktur gesorgt, die sowohl das Erstellen als auch Löschen von Daten vorsieht. Da diese im Laufe der Zeit eine unterschiedliche Verfügbarkeitsnotwendigkeit erhalten, wird für Automatismen gesorgt, die die Informationen optimal in einer hybriden Speicherlandschaft ablegen: Zum Beispiel haben kaufmännische Daten eines Unternehmens eine hohe Zugriffs- sowie Veränderungswahrscheinlichkeit während des aktuellen Geschäftsjahres. Nach Buchhaltungsschluss sind keinen Modifikationen erlaubt, dem entsprechend sinkt der Zugriff. Werden solche Daten aus dem hochperformanten Bereich entfernt, entlastet dies nicht nur die Speichermedien, sondern auch die Verwaltungs- und Rechenkapazitäten.

Aber ILM behandelt die Datenverwaltung nicht nur aus Sicht der IT-Infrastruktur, sondern auch aus der Management-Perspektive, die „Policy Management und Dienstgütevereinbarungen“ festlegt, welche sich am Geschäftsmodell orientieren.

Hierbei werden Richtlinien für die Informationsverwaltung extrahiert, welche dann für die Klassifizierung des Wertes zuständig sind. Wichtige Voraussetzung für die Umsetzung sind klar definierte und funktionierende Prozesse sowie angemessene Infrastrukturen, die in einer engen Verzahnung aufeinander aufsetzten können.

An Hand von Metadaten werden Informationen im Kontext des Geschäftsmodells interpretierbar und können somit in verschiedene Wertigkeitsklassen eingeteilt und einem kosteneffizienten Speichermedium zugeordnet werden, ohne die Dienstgütevereinbarungen zu verletzen.

Ebenfalls im Konzept sind die Menschen integriert, die mit den Daten operieren („Applikationen, Metadaten"). Dabei muss sichergestellt werden, dass zugängliche Richtlinien für den Einzelanwender vorhanden sind, deren Einhaltung mittels eingeführter Kontrollinstanzen sichergestellt wird. Somit verbindet ILM die Speicherinfrastruktur mit dem bestehenden Geschäftsmodell und berücksichtigt ebenfalls die Menschen innerhalb der Prozesse.

Der zweite Teil der Definition betrachtet ILM mit Blick auf die Zukunft, da es sich um eine Strategie handelt, die mit den Prozessen mitwachsen soll:

> *ILM-Vision: Eine neue Menge von Management-Praktiken, die den Wert einer Information innerhalb eines Geschäftsmodells auf eine adäquate und kosteneffizente Infrastruktur abbilden.*
>
> *Quelle:[1], Eigene Übersetzung*

Besonders hervorzuheben sind hier die „neuen Mengen von Management-Praktiken", da die heute angewandten Prozesse in der Speicherverwaltung den aufkommenden Anforderungen der Zukunft nicht gerecht werden.

In den Datenzentren existieren bereits lokale Lösungen, die zumeist darauf ausgerichtet sind, eine Teilmenge des Volumens adäquat und kosteneffizient zu verwalten. Die Herausforderung von ILM besteht nun darin, diese Teillösungen zu lokalisieren, um sie iterativ zu einer Gesamtlösung zusammenzufügen.

2.3 Notwendigkeit der Betrachtung

Das IT-Speicherumfeld hat sich in den letzen Jahren erheblich verändert. Die Verwaltung von Daten erfordert professionelle und umfassende Konzepte, um den internen und externen Vorgaben in Form von gesetzlichen Bestimmungen und Verordnungen sowie zunehmendem Kostendruck und technologischen Innovationen gerecht zu werden. Nachfolgend werden einzelne Faktoren erläutert:

- Datenzuwachs
- Gesetzliche Bestimmungen
- Variierender Wert der Daten
- Kostendruck
- Technologische Innovation

2.3.1 Datenzuwachs

Jedes Unternehmen beziehungsweise jede Institution kann ein steigendes Datenvolumen in den unterschiedlichsten Bereichen verzeichnen.

Ein Beispiel für extrem hohen Datenzuwachs findet sich im Gesundheitswesen. Ein Picture Archive and Communications Systems (PACS) erzeugt in einem größeren Krankenhaus mehrere Terabytes pro Jahr an digitalen Röntgenbildern und Kernspintomographien. In absehbarer Zukunft müssen diese über die Lebensdauer eines Pa-

tienten hinaus erhalten werden, was enorme Archivierungskosten verursachen wird. Neben den reinen Hardwarekosten müssen Verwaltungs- sowie Verwahrungsaufwand berücksichtigt werden, was zu weiteren Personal- und Raumkosten führen wird.

> *"For every $1 spent on storage hardware, $3,50 is being spent on storage administration."*
>
> *Ray Paquet, Gartner Group, April 2002*

2.3.2 Gesetzliche Bestimmungen und Verordnungen

Gesetzliche Bestimmungen und Verordnungen legen die Form und Dauer der Datenarchivierung fest und bestimmen, welche Dienstgütekriterien eingehalten werden müssen. Die Basis sind nationale und internationale Vorgaben, die meist von internen Firmen- und Institutionsvorgaben ergänzt werden.

In Deutschland greifen für ein Unternehmen die Grundsätze der ordnungsgemäßen Buchführung (GoB), die Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GPDdU), das Handelsgesetzbuch (§ 238 HGB) und die Abgabenordnung (§§140-148 AO), die verschiedene Fristen und Formen der Datenverwahrung vorsehen.

Je nach Geschäftsbranche gibt es weitere Bestimmungen, wie zum Beispiel die Vorgaben der Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), die für Banken, Versicherungen sowie für den Aktienhandel relevant sind. Insgesamt ist eine Gesetzgebung zu erwarten, die den Umfang der ordnungsgemäß zu speichernden Informationen und die Länge der Aufbewahrungsfristen anheben wird.

2.3.3 Variierender Wert der Daten

Der Wert von Daten leitet sich von seiner Zugriffswahrscheinlichkeit ab. Eine Studie von Horison Information Strategies [5] besagt, dass nach drei zugriffsfreien Tagen die Wahrscheinlichkeit einer Wiederverwendung von Office-Dokumenten bei 50 Prozent liegt, nach 30 Tagen nur mehr bei wenigen Prozent.

Wie sich im Laufe der Zeit die Wertigkeit von Daten verändert, soll exemplarisch mittels eines vereinfachten Banktransaktions-Beispiels vorgestellt werden: Die wichtigste Information ist der Kontostand, der eine hohe Zugriffswahrscheinlichkeit hat, da sich daraus Forderungen und Verbindlichkeiten ableiten. Eine konkrete Umsatzbewegung (Abhebung oder Einzahlung) wird so lange auf dem Speicherbereich mit dem schnellsten Dateizugriff gehalten, bis der Kontoauszug abgerufen wird. Ab diesem Zeitpunkt werden einzelne Einträge nur für die Berechnung der Umsätze sowie der Bilanzsumme benötigt und dementsprechend in einem kostengünstigeren Speichermedium eingelagert.

Nach Veröffentlichung der Bilanzsumme spielen sämtliche Einzelwerte eine untergeordnete Rolle und werden nur in Fällen von Reklamation oder Steuerprüfung benötigt, was zu einer geringen Anzahl von Zugriffen führt. Nach Ablauf der Reklamationsfrist beziehungsweise der vorgeschriebenen Archivierungszeit liegt es im Ermessen der Bank, ob und für welche Dauer die Daten vorgehalten werden.

Daraus lässt sich erkennen, dass eine Aufteilung der Speicherlandschaft in verschiedene Hierarchien mit unterschiedlichen Kosten sinnvoll ist, weil eine kostengünstigere Verwahrung vorgenommen werden kann, ohne die Informationsverfügbarkeit für den Anwender einzuschränken.

Anhand eines Datenbestandes eines DAX-30-Unternehmens wird nun evaluiert, ob die Ergebnisse von Horison Information Strategies zutreffen, um auf deren Basis ein ILM Konzept aufzusetzen.

2.3.4 Kostendruck

Zwar sind die Prozesse und deren Kosten in den Unternehmen sehr individuell, jedoch bringt eine Automatisierung in den meisten Fällen eine erhebliche Ressourcenreduktion und somit auch eine Einsparung.

Da ILM automatisierte Prozesse aufbauend auf vorhandenen Technologien und Ressourcen vorsieht, die optimal ausgenutzt werden, ergibt sich dadurch eine Reduzierung von Personal- und Raumkosten ohne Zugriffseinschränkungen. Somit erzielt die Umsetzung von ILM langfristig Gewinne, da trotz wachsendem Datenvolumen die Speicherinfrastruktur durch Entsorgung irrelevanter Informationen konstant gehalten werden kann.

2.3.5 Technologische Innovationen

Das Aufkommen neuer Speichertechnologien, wie zum Beispiel ATA[1]-Platten, mit gestaffelten Preisen und Leistungen, sowie Fibre Channel und SCSI[2]-Platten erweitern das klassische Speichermodell. Zurzeit geht man von einem dreistufigen Modell aus, wenn die verschiedenen Technologien im Online- und Nearline-Bereich zusammen betrachtet werden.

An der Spitze steht der hochperformante Online-Bereich, dessen wesentliche Charakteristik die unverzügliche Verfügbarkeit von Daten ist, mit Ein- und Ausgabezeiten im Bereich von wenigen Millisekunden. Im Anschluss daran befinden sich Systeme mit ATA-Platten, die auch eine direkte Verfügbarkeit gewährleisten, allerdings mit moderateren Durchsatzwerten und geringeren Kosten pro Megabyte. Diese werden hauptsächlich zur Archivierung und in der Platte-zu-Platte Datensicherung eingesetzt.

An der Basis des Modells steht der Nearline-Bereich mit den Bandtechnologien, der sich durch geringe Kosten pro Megabyte und hohe Zugriffszeiten auszeichnet. Werden die Bänder automatisch in Bibliotheken verwaltet, kann mit Ein- und Ausgabezeiten im Sekunden- und Minutenbereich gerechnet werden. Sind diese jedoch ausgelagert, muss aufgrund von manueller Bereitstellung von noch höheren Zeiten ausgegangen werden.

Eine Herausforderung für ILM besteht darin, die verschiedenen Dienstgütekriterien der Daten zu erkennen und diese der adäquaten Speicherschicht zuzuordnen.

2.4 Umsetzungsstrategie

Für die Umsetzung von ILM werden fünf Phasen [4] vorgeschlagen, die schrittweise das Framework in ein konkretes Geschäftsmodell einbinden:

- Erfassung: Die Speicherlandschaft des Unternehmens wird mittels Verwaltungssystemen untersucht, um die Daten auf den Speichermedien zu erfassen.
- Sozialisierung: Die Ergebnisse der Ist-Analyse werden der Unternehmensleitung präsentiert und gemeinsam die Verwahrungsorte der Dateien mit den Geschäftsprozessen abgeglichen.

[1] Advanced Technology Attachments

[2] Small Computer System Interface

- Klassifizierung: Den Informationen werden Werte zugewiesen. Dazu gibt es verschiedene Möglichkeiten. Das konkrete Vorgehen basiert auf dem Ergebnis der Phase Sozialisierung.
- Automatisierung: Eine Optimierungsfunktion verteilt die Informationen entsprechend ihrer Klassifikation auf ein adäquates Speichermedium unter Berücksichtigung der Dienstgütevereinbarungen.
- Überprüfung: Nach Einführung von Automatisierungsprozessen wird in regelmäßigen Abständen überprüft, ob die Abbildung des Geschäftsmodells noch gültig ist, da dieses sich über den Zeitraum verändern kann, was dann in ILM eingearbeitet werden muss.

Als Teil der Erfassungsphase ist das Potential einer ILM-Lösung zu identifizieren. Dazu verschafft man sich einen Gesamtüberblick über die Speicherlandschaft und der Daten.

Bei den Daten untersucht man nach Zugriffe. Aus erkannten Mustern lassen sich Regeln zur Migration erstellen, die in der Phase 3 zur Verwendung kommen.

Nachfolgende Untersuchungen der Dateizugriffe dient der Ermittlung des Potentials in dem konkreten Fall einer realen Datenbank.

2.5 Vergleichbare Untersuchungen

Untersuchungen über Zugriffe auf Dateien wurden u.a. von Strange et al. 1992 [7] vorgenommen. Ebenso haben sich Gibson et al 1998 [8] mit Zugriffen auf Dateien betrachtet. Die zu Grunde liegenden Zeiträume lagen zwischen 120-280 Tagen. Die betrachteten Dateien gehörten zu verschiedenen Einheiten von Universitäten und der Armee in den USA.

Zugriffe auf Informationsobjekte werden häufig mit der sogenannten Zipf-Verteilung modelliert. Studien haben gezeigt, dass Napster-, Gnutella- und Internetabfragen dieser Verteilung folgen [9]. Es gibt allerdings auch gegenteilige Erkenntnisse [10].
Eine weitere Studie, die explizit im Umfeld von ILM entstand, ist die Studie von Horison Information Strategies aus den Jahr 2004 [5]. Diese wurde im Auftrag durch die Universität Berkeley erstellt und beschränkte sich auf einen beobachteten Zeitrahmen von 90 Tagen.
Die Hauptaussagen der Studie lauten:

1. Die Wahrscheinlichkeit, Daten drei Tage nach Erstellung wieder zu verwenden, fällt auf 50 Prozent.
2. Die Wahrscheinlichkeit, Daten dreißig Tage nach Erstellung wieder zu verwenden, fällt auf unter wenige Prozent.[3]

[3]Wenige Prozent wird nachfolgend mit <= 5 Prozent interpretiert.

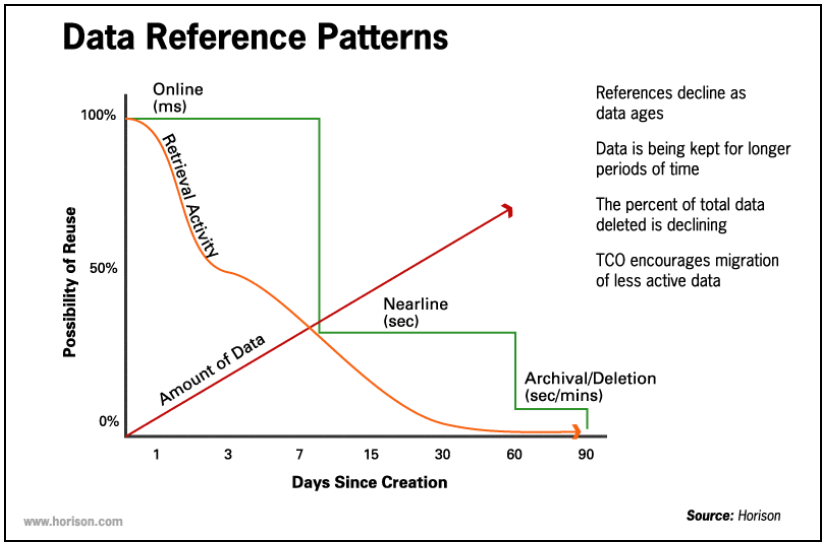

Abbildung 1: Visualisierte Ergebnisse der Horison-Studie [5]

Die Abgrenzung der vorliegenden Studie zu den bekannten Studien offenbart sich darin, dass es sich hier um eine Windows-Umgebung mit einem aktuellen Datenbestand handelt. Ferner wird nur eine Unternehmens-Einheit untersucht, aus der mehrere Projekte analysiert werden. Der Datenbestand entstammt einem DAX-Unternehmen und resultiert aus verschiedenen Applikationen.

Es werden die Hauptaussagen der Horison-Studie als Kriterium für die eigenen Untersuchungen herangezogen. Dazu werden zuerst die relativen Zugriffshäufigkeit der verschiedenen Dokumente ermittelt.

3 Ausgangssituation

Im Rahmen dieses Technical Reports wurde eine Analyse über ein Dokumentenverwaltungssystem eines DAX-30-Unternehmens vorgenommen, um zu evaluieren, ob genügend Potenzial für die Umsetzung von ILM vorhanden ist.

Das Dokumentenverwaltungssystem wird von einem deutschlandweit tätigen Consulting-Bereich benutzt. Dieser Bereich unterteilt sich in neun Regionen, die jeweils ein eigenes Consulting Team haben, welches in der Region Rhein-Main etwa 90 Mitarbeiter zählt; deutschlandweit sind es circa 700 Personen.

Das Unternehmen unterstützt die überregionale Zusammenarbeit der einzelnen Consulting-Teams. Hierfür wurde ein einheitliches Dokumentenverwaltungssystem, nachfolgend „Datenbank“ genannt, eingeführt, in der sämtliche Berater ihre Projekte ablegen müssen. Damit wird ein überregionaler Zugang zu allen Dokumenten gewährleistet.

Die gesamte Datenbank umfasst über 150.000 Dokumente. Die Dokumente können in der Datenbank editiert, angesehen, versioniert und gelöscht werden.

Beim Versionieren werden mehrere physikalische Dateien für ein Dokument im System vorgehalten. Ein Löschvorgang macht diese nicht mehr über die üblichen Clients zugänglich, obwohl sie drei Tage in der Datenbank vorgehalten werden und nur mit einem Administratorzugang wieder hergestellt werden können.

Eine Untergruppe, das Projektmanagement (PM), hat klare Richtlinien für die Dokumentation und Ablage, welche für eine einheitliche Archivierung sorgen sollen.

Aus diesem Grund wurden Stichproben vom PM-Bereich entnommen, da definierte Vorgaben und Prozesse bestehen und abgeglichen werden können.

Ziel der Studienarbeit ist es zu identifizieren, wie häufig Dateien aufgerufen werden. Abgeleitet daraus sollen nicht nachgefragte Dateien als Potential für ILM identifiziert werden.

4 Die gesammelten Daten

Am 1.11.2004 wurden in der Datenbank 134 Projekte gezählt. Die Datenbank selbst unterteilt sich in drei Bereiche, benannt Mittelwest (MW), Nordost (NO) sowie Süd-Südwest (SS), in denen die neun Regionen ihre Projektablage vornehmen. Um keinen Bereich zu bevorteilen, steuerte jeder Bereich den jeweils gleichen Anteil an der Stichprobe bei, da keine Mitarbeiterverteilung bekannt war.

Die alphabetische Liste der Projekte wurde durchnummeriert. Die vergebenen Ordnungsnummern identifizierten die Projekte. Mittels der Random-Funktion von Java wurde eine Pseudozufallszahl im Bereich [0, n-1] (n = Anzahl Projekte) bestimmt. Nach der Auswahl wurde dieses aus der Liste gestrichen und selbige neu durchnummeriert.

Dieser Schritt wurde für jeden Bereich dreimal ausgeführt mit dem Ergebnis einer neunelementigen Stichprobe. Die Auswahlwahrscheinlichkeit für jedes Element der Grundgesamtheit war somit gleich, worauf es hier zunächst ankommt.

Daher wurden 1762 Einzeldokumente, die insgesamt 942 MB Speicher benötigen, protokolliert. Aus Gründen der Anonymisierung wurden die untersuchten Projekte mit Großbuchstaben beschriftet, mit denen sie nachfolgend referenziert werden (siehe Tabelle 1).

Projekt	A	B	C	D	E	F	G	H	I
Region	MW	MW	MW	NO	NO	NO	SS	SS	SS
Dateien	196	170	121	430	592	6	116	66	65
Menge (KB)	54687	267840	26725	338295	191848	640	36483	12952	12969

Tabelle 1: Untersuchte Projekte

Der protokollierte Zeitraum eines Dokumentes erstreckt sich von der Erstellung bis zum 31. Januar 2005 und wurde über die Notification[4] sowie zur Konsistenzprüfung über die History der Einzeldokumente bestimmt. Aus diesen Rohdaten wurde eine relative Zugriffshäufigkeit für jedes Intervall berechnet.

Die Einteilung der zu betrachtenden Zeitintervalle erfolgt in Anlehnung an die Horison–Studie und lautet:

(0, 1], (1, 3), [3, 7), [7, 15), [15, 30), [30, 60), [60, 90), [90, ∞[5]) Tage.

In Kapitel 7 wird der Zeitraum jenseits der 90 Tage weiter unterteilt und untersucht, weil sich zeigen wird, dass eine Aussage jenseits von 90 Tagen einer weiteren Differenzierung bedarf.

[4]Logbuch der Zustandsveränderung eines Dokumentes in der Datenbank

[5]Exakt endet das letzte Intervall mit dem Tag der Datenerhebung.

5 Auswerten der Zugriffsdaten auf 90-Tage-Basis

Die einzelnen Dokumente der Datenbank werden untersucht. Die Einteilung der zu betrachtenden Zeitintervalle lautet:

(0, 1], (1, 3), [3, 7), [7, 15), [15, 30), [30, 60), [60, 90), [90, ∞[6]) Tage.

Um die relativen Zugriffe auf die vorgegebenen Intervalle zu erhalten, wird wie folgt vorgegangen: Es seien D1, D2, D3 drei Dokumente, die folgende Zugriffe aufweisen (Tabelle 2, links). Findet für ein Intervall und ein Dokument ein Zugriff statt, so wird eine Eins notiert, ansonsten eine Null (Tabelle 2, rechts). In jedem Intervall werden die Zugriffe summiert und durch die Anzahl der Dokumente dividiert.

Zugriffe	Intervall 1	Intervall 2	Intervall 3
D1	12	4	2
D2	10	3	0
D3	9	0	0

Zugriffe	Intervall 1	Intervall 2	Intervall 3
D1	1	1	1
D2	1	1	0
D3	1	0	0
Relative Z.	1	0,66	0,33

Tabelle 2: Erhebung Nutzdaten

Diese Schritte werden für alle untersuchten Dokumente und Intervalle vorgenommen.

5.1 Aggregierte Auswertung

Die aggregierte Auswertung führt zu folgender relativer Zugriffshäufigkeit (siehe Tabelle 3 und Abbildung 3).

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0,82	6,59	2,35	8,74	4,1	15,37	9,61
Zugriff abs.	3909	52	151	134	408	142	685	397,58

Tabelle 3: Relative Zugriffshäufigkeiten der Dokumente

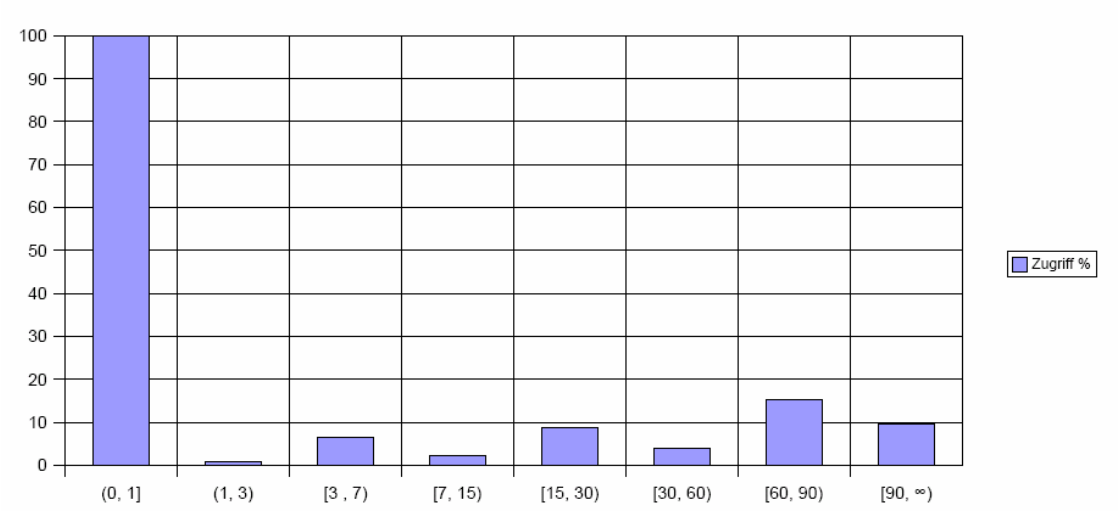

Abbildung 2: Relative Zugriffshäufigkeiten der Dokumente

Die Entwicklung der Zugriffe weist ein deutliches Gefälle auf. Die Zugriffshäufigkeitn im Intervall drei bis sieben Tage liegt bei 6,59 Prozent. Dies entspricht in etwa der ersten Hauptaussage der Horison-Studie.

[6]Exakt endet das letzte Intervall mit dem Tag der Datenerhebung und ist somit endlich.

Die Zugriffshäufigkeit im Intervall [30, 60) liegt unter 5 Prozent, nämlich bei 4,61 Prozent. Werden die beiden nächsten Intervalle untersucht, ist ein Zuwachs ersichtlich, nämlich 15,37 Prozent im Intervall [60, 90) und 9,61 Prozent im Intervall [90, ∞).

Somit kann die zweite Hauptaussage der Horison-Studie, dass nach 90 Tagen die Zugriffshäufigkeiten nahe Null liegen, hier nicht bestätigt werden. Dies bedeutet, dass die Horison-Studie allein nicht als Grundlage für ILM bei der Datenbank herangezogen werden sollte. Um exaktere Aussagen über die Zugriffe machen zu können, werden alle neun Projekte der Stichprobe einzeln untersucht.

5.2 Auswertung der einzelnen Projekte auf 90-Tage-Basis

Die einzelnen Projekte wurden aus Gründen der Anonymisierung mit den Buchstaben A bis I versehen.

5.2.1 Projekt A

Das Projekt A beinhaltet 196 Dokumente, die mit 54687 MB 5,8 Prozent des untersuchten Datenvolumens darstellen. Wie aus Tabelle 4 und Abbildung 4 ersichtlich, trifft die erste Aussage der Horison-Studie zu, da die Zugriffshäufigkeit nach drei Tagen bei circa 50 Prozent liegt. Auch die zweite Aussage kann bestätigt werden, da die Zugriffshäufigkeit nach 30 Tagen Null Prozent hat. Werden die beiden darauf folgenden Intervalle näher untersucht, sind noch Zugriffe zu vermerken, jedoch liegen diese bei 2,55 Prozent der Projektdokumente. Somit haben über 97 Prozent keinen Zugriff 90 Tage nach ihrer Erstellung.

Die Zugriffe sind sehr unharmonisch verteilt. Das kann daran liegen, z. B. dass das Projekt zum Zeitpunkt der Einstellung in die Datenbank bereits fortgeschritten war und in der Untersuchung nur noch das Ende des Projektes in den Zugriffen wahrgenommen wird.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0	51,02	0	0	0	0	2,55
Zugriff abs.	196	0	100	0	0	0	0	5

Tabelle 4: Projekt A - Relative Zugriffshäufigkeiten

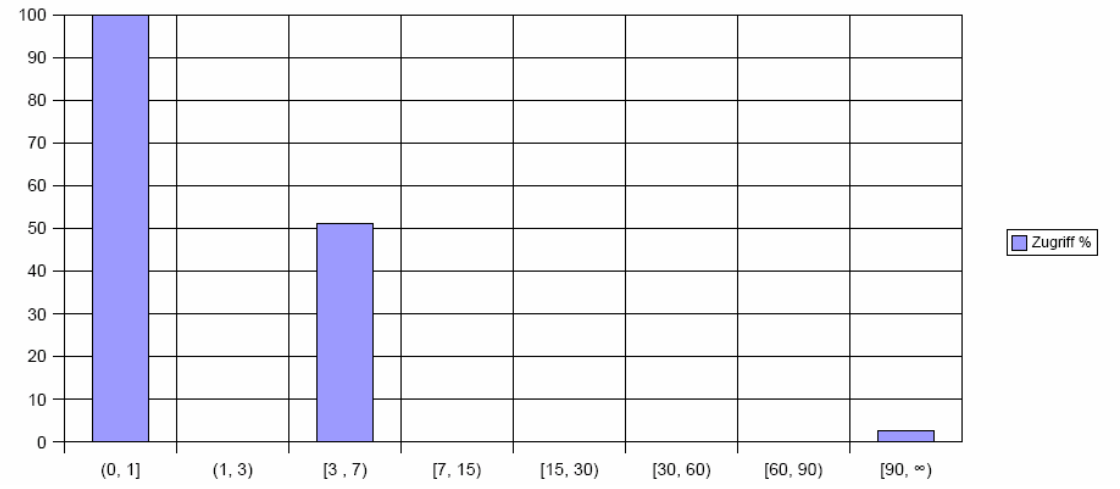

Abbildung 3: Projekt A - Relative Zugriffshäufigkeiten

5.2.2 Projekt B

Das Projekt B (Tabelle 5 und Abbildung 5) beinhaltet 170 Dokumente, die mit 267840 MB 28,4 Prozent des untersuchten Volumens darstellen. 99,41 Prozent der Projektdokumente haben 90 Tage nach ihrer Erstellung keinen Zugriff mehr zu verzeichnen.

Grund für die extreme Verteilung der zugriffe kann sein, dass z.B. das Projekt erst nach Abschluss in die Datenbank eingepflegt wurde und somit bereits bei Einstellung abgeschlossen war.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0	0	0	0	0	0	0,59
Zugriff abs.	170	0	0	0	0	0	0	1

Tabelle 5: Projekt B - Relative Zugriffshäufigkeiten

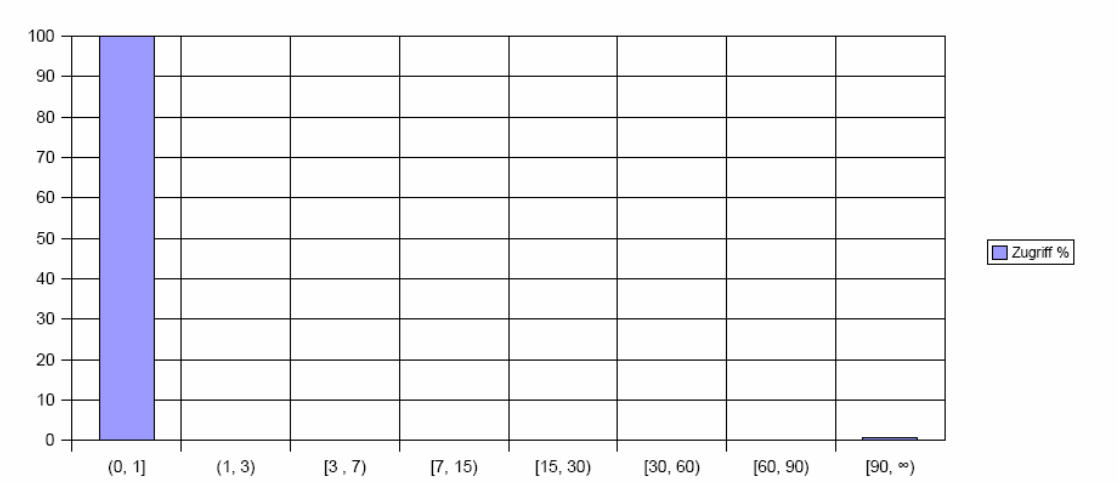

Abbildung 4: Projekt B - Relative Zugriffshäufigkeiten

5.2.3 Projekt C

Das Projekt C umfasst 121 Dokumente mit insgesamt 26725 MB (2,8 Prozent des untersuchten Volumens). Tabelle 6 und Abbildung 6 zeigen, dass dieses Projekt die Aussage der Horison-Studie nicht vollständig bestätigen kann, da die Häufigkeit nach 30 Tagen bei 14,05 Prozent liegt und auch nicht mehr signifikant sinkt. Immerhin haben über 86 Prozent keinen Zugriff 90 Tage nach ihrer Erstellung.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0,83	0,83	4,13	4,96	14,05	13,22	13,22
Zugriff abs.	140	1	1	5	6	17	16	16

Tabelle 6 : Projekt C - Relative Zugriffshäufigkeiten

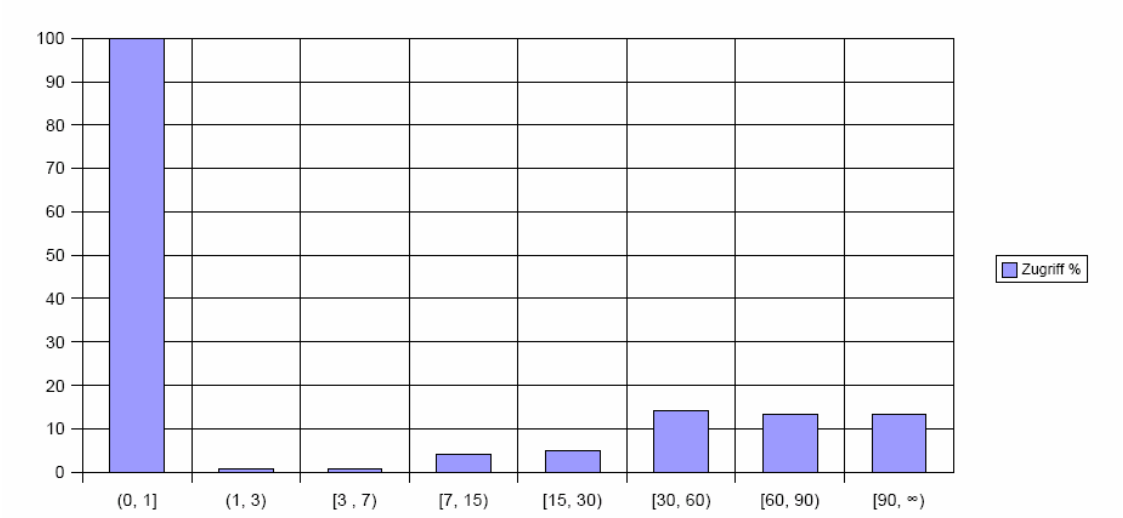

Abbildung 5: Projekt C - Relative Zugriffshäufigkeiten

5.2.4 Projekt D

Im Projekt D (Tabelle 7 und Abbildung 7) befinden sich 430 Dokumente (338295 MB entsprechend 35,8 Prozent des untersuchten Volumens). In den ersten beiden Wochen nach Erstellung verzeichnet dieses Projekt ähnlich wenige Zugriffe wie A bis C, während auffällige Spitzen in den Intervallen [15, 30), [60, 90) auftreten. Bei diesem Projekt sind während der Abwicklung Probleme aufgetreten, was die überdurchschnittlichen Zugriffsraten erklärt. Nimmt man die beiden Intervalle aus der Betrachtung heraus, erkennt man eine immer noch steigende Zugriffstendenz, die im Widerspruch zur Horison-Studie steht.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	1,63	1,4	3,95	42,09	4,42	65,12	6,28
Zugriff abs.	1437	13	6	34	190	23	304	5,58

Tabelle 7: Projekt D - Relative Zugriffshäufigkeiten

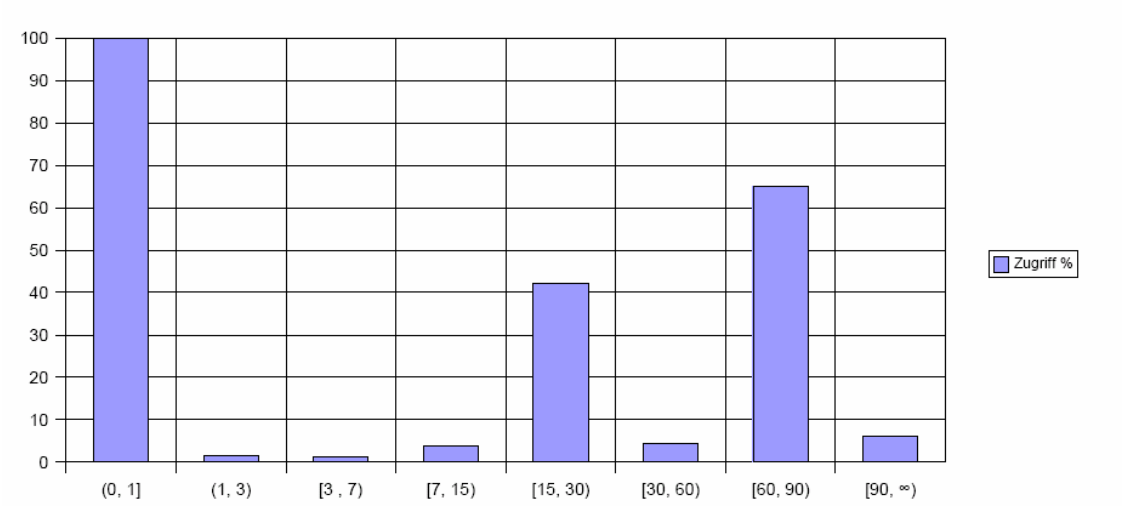

Abbildung 6: Projekt D - Relative Zugriffshäufigkeiten

5.2.5 Projekt E

Das Projekt E beinhaltet 592 Dokumente, die mit 191484 MB 20,3 Prozent des untersuchten Volumens darstellen. Wie aus Tabelle 8 und Abbildung 8 ersichtlich, verletzen die späten Zugriffe auf die Projektdokumente die Aussagen der Horison-Studie deutlich: Nach 30 Tagen wächst die Zugriffsrate von 10,64 Prozent auf 54,22 Prozent im Intervall [60, 90) und sogar auf 62,16 Prozent in [90, ∞)

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	4,9	5,24	9,63	29,9	10,64	54,22	62,16
Zugriff abs.	1709	38	42	86	208	91	359	368

Tabelle 8: Projekt E - Relative Zugriffshäufigkeiten

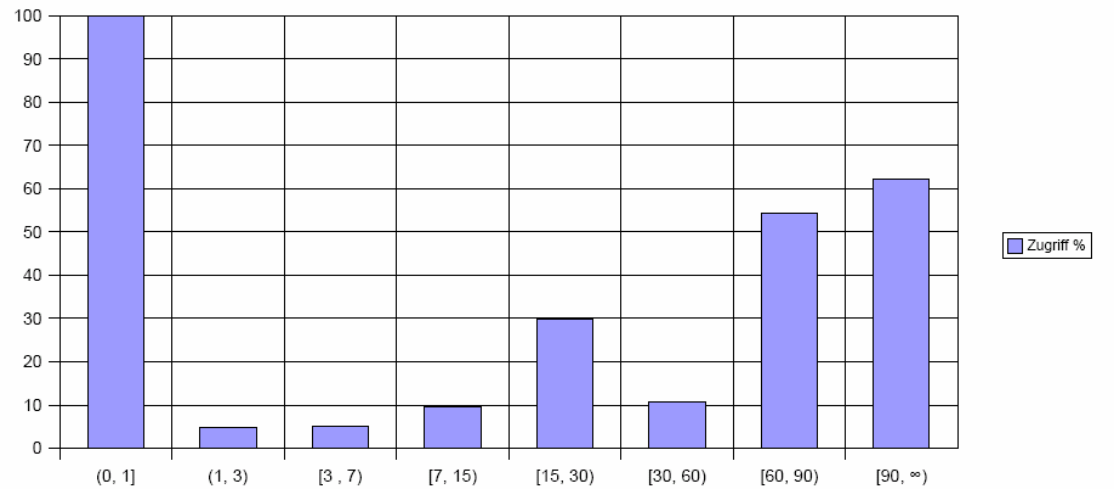

Abbildung 7: Projekt E - Relative Zugriffshäufigkeiten

5.2.6 Projekt F

Das Datenvolumen von Projekt F ist mit 6 Dokumenten (640 MB, 0,0006 Prozent des untersuchten Volumens) deutlich kleiner als die anderen Elemente der Stichprobe Tabelle 9 und Abbildung 9).

Ebenso wie bei Projekt B sind keine Zugriffe nach Einstellung in die Datenbank zu verzeichnen. Grund für die extreme Verteilung der Zugriffe kann sein, dass z.B. das Projekt erst nach Abschluss in die Datenbank eingepflegt wurde und somit bereits bei Einstellung abgeschlossen war.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0	0	0	0	0	0	0
Zugriff abs.	6	0	0	0	0	0	0	0

Tabelle 9: Projekt F - Relative Zugriffshäufigkeiten

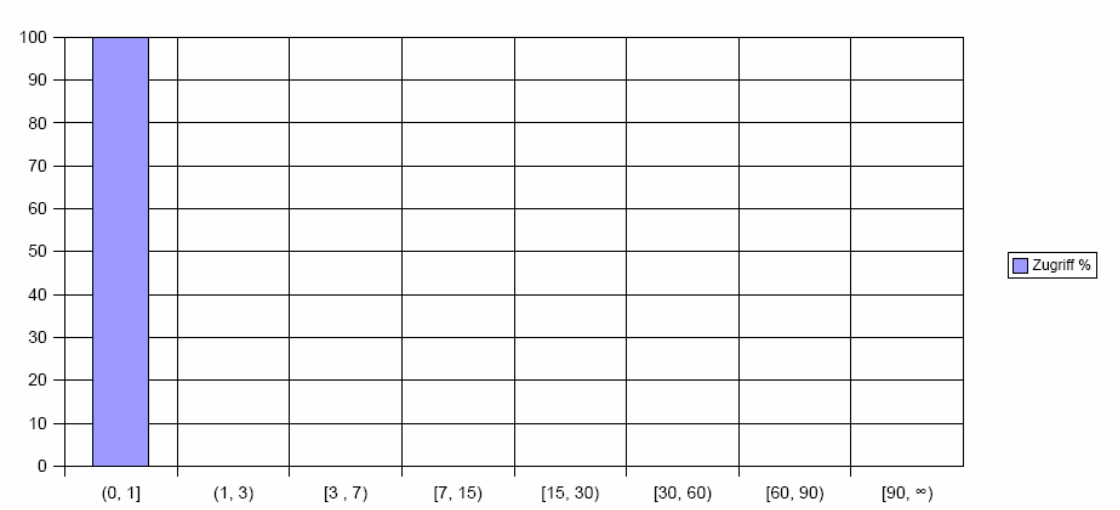

Abbildung 8: Projekt F - Relative Zugriffshäufigkeiten

5.2.7 Projekt G

Projekt G (Tabelle 10 und Abbildung 10) ist mit 116 Dokumenten, die mit 36483 MB 3,8 Prozent des untersuchten Volumens darstellen, im Vergleich zu Projekt F ein typischeres Projekt. Nach Erstellung wird auf die meisten Dokumente gar nicht mehr und auf alle anderen nur noch sehr selten zugegriffen.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0	0,86	3,45	1,72	7,76	2,59	1,72
Zugriff abs.	119	0	2	9	4	11	6	2

Tabelle 10: Projekt G - Relative Zugriffshäufigkeiten

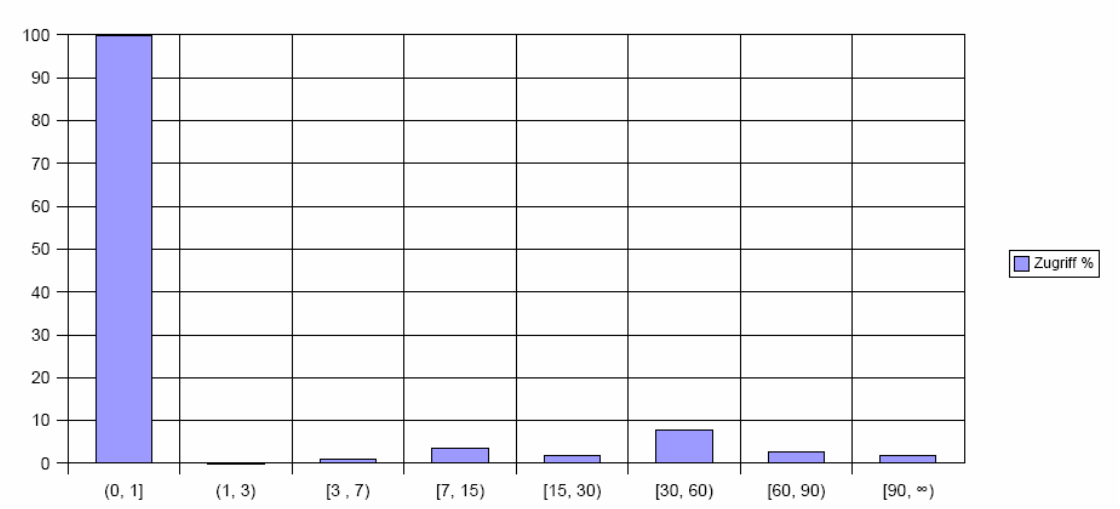

Abbildung 9: Projekt G - Relative Zugriffshäufigkeiten

5.2.8 Projekt H

Das Projekt H (Tabelle 11 und Abbildung 11) ähnelt in der Zugriffsstatistik sehr dem Projekt F insofern, als nach den ersten beiden Tagen überhaupt keine weiteren Zugriffe auf den Projektdokumenten registriert wurden. Bemerkenswert dabei ist, dass beide Projekte größenordnungsmäßig nicht vergleichbar sind (66 Dokumente, 12952 MB, 1,37 Prozent des untersuchten Volumens).

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0	0	0	0	0	0	0
Zugriff abs.	67	0	0	0	0	0	0	0

Tabelle 11: Projekt H - Relative Zugriffshäufigkeiten

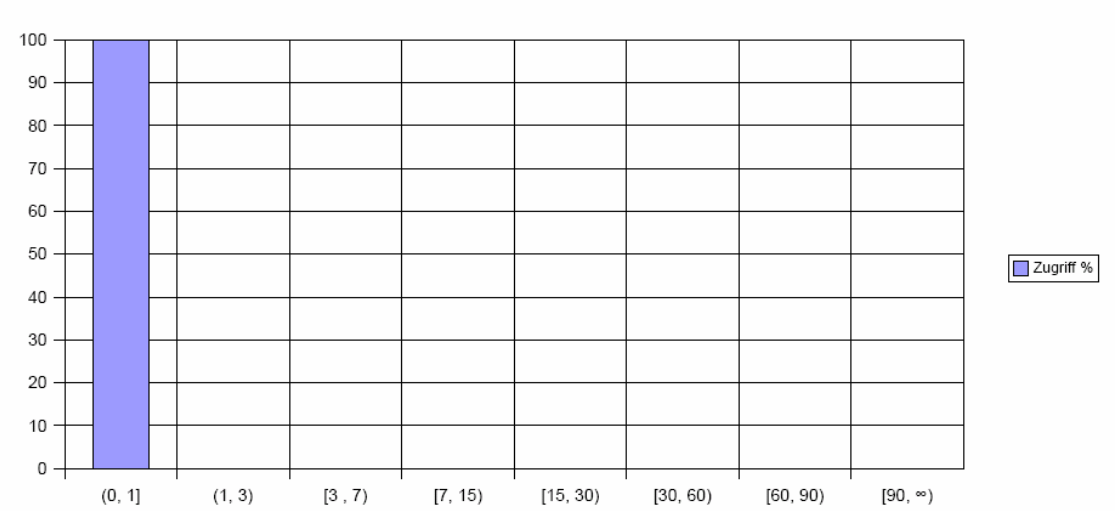

Abbildung 10: Projekt H - Relative Zugriffshäufigkeiten

5.2.9 Projekt I

Für das Projekt I (Tabelle 12 und Abbildung 12) gelten dieselben Aussagen wie für Projekt H .Es beinhaltet 65 Dokumente, die mit 12969 MB 1,37 Prozent des untersuchten Volumens darstellen.

Intervall	(0, 1]	(1, 3)	[3, 7)	[7, 15)	[15, 30)	[30, 60)	[60, 90)	[90, ∞)
Zugriff %	100	0	0	0	0	0	0	0
Zugriff abs.	65	0	0	0	0	0	0	0

Tabelle 12: Projekt I - Relative Zugriffshäufigkeiten

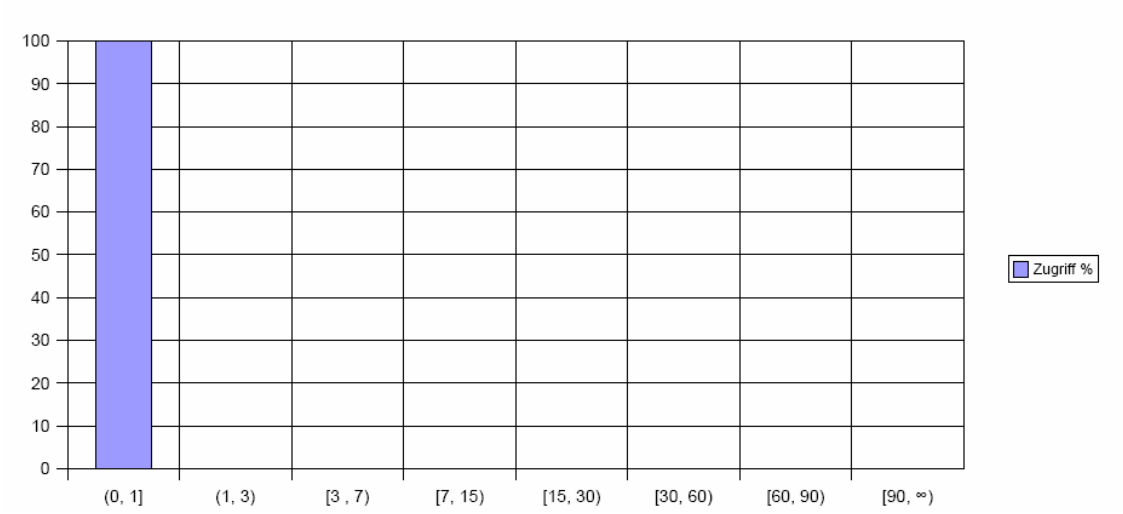

Abbildung 11: Projekt I - Relative Zugriffshäufigkeiten

5.3 Ergebnis der Auswertungen auf 90-Tage-Basis

Die Einzelbetrachtung der Projekte zeigt, dass entgegen den Aussagen der Horison-Studie 30 Tage nach Erstellung eines Dokumentes und später sehr wohl noch Zugriffe in signifikanter Häufigkeit stattfinden.

Auch wenn man über alle Projekte aggregiert, sinkt die Zugriffshäufigkeit erst 90 Tage nach Erstellung unter die 10-Prozent-Schwelle.

Wie man an der großen Streuung später Zugriffshäufigkeiten zwischen den einzelnen Projekten festgestellt hat, stellt sich grundsätzlich die Frage, ob eine solche Aggregierung sinnvoll ist.

Diese Frage ließe sich mit statistischen Mitteln beantworten, indem man z. B. auf Anpassung an die „Horison-Kurve“ testet. Da die empirische Verteilung der Horison-Kurve nicht vorliegt, ist eine Anpassung ausgeschlossen. Angesichts dessen, dass die Horison-Studie schon mittels deskriptiver Statistik als nicht anwendbar erkannt wurde, empfiehlt sich eine Anpassung an andere bekannte Verteilungen.

6 Auswertung der Dokumentypen der einzelnen Projekte

Für jedes Projekt wird die Aufteilung des Speicherbedarfs und die Anzahl der Dokumente jeweils in Abhängigkeit vom Dokumententyp näher betrachtet. Anhang B enthält eine kurze Erklärung, von welcher Applikation dieser stammt.

Die diversen Dokumentkategorien (wie Datenhaltung [MDB, LDB, XLS, MPP], Datenaufbereitung [DOC, PPT, VSS, VSD], Archivierung [PDF, GZ, ZIP], Illustration [TIF, JPG, GIF] sowie Sonstige [XLA, DOT, MSG, TXT, HTM, RTF, TRC]) unterscheiden sich hinsichtlich Anzahl zugeordneter Dokumente und deren Größe teilweise beträchtlich.

Nach Betrachtung der einzelnen Projekte wird nun ein Vergleich mit den aggregierten Daten vorgenommen. Stellt sich heraus, dass die Ergebnisse im Wesentlichen projektunabhängig ausfallen, würde dies ein einfacheres Regelkonzept für ILM erlauben, ohne dass die Effektivität des Konzepts merklich leiden würde.

Zu diesem Zweck weisen die nachfolgenden Statistiken wie bisher die Abhängigkeit von Dokumentenzahl beziehungsweise -größe nach Dokumententyp aus, jedoch ohne Berücksichtigung der Projektzugehörigkeit. Um die Repräsentativität dieser Ergebnisse einzuschätzen, wird zusätzlich die empirische Streuung bezüglich Projektzugehörigkeit und Dokumentenkategorien ermittelt.

Die Kategorien Illustration und Sonstige tragen zur Gesamtzahl der Dokumente mit 2,64 Prozent einen verschwindend geringen Teil bei, weswegen sie nicht weiter diskutiert werden.

Anhand von Tabelle 13 und 14 und Abbildung 13 und 14 ergeben sich für die Dokumentenkategorien folgende Anteile hinsichtlich Gesamtdokumentzahl und Gesamtspeicherbedarf (Tabelle 15):

Projekte	Aufbereitung	Haltung	Archivierung
Anzahl Dokumente	850	447	401
Anteil Dokumentenzahl	48,24	25,37	22,76
Speicherbedarf (KB)	556864	106604	256801
Anteil Speicherbedarf	59,12	11,32	27,27

Tabelle 13: Dokumentenanzahl und Speicherbedarf nach Kategorien

Die Aufschlüsselung nach Dokumententypen ähnelt bezüglich der Dokumentenzahl stark dem Projekt E. Hinsichtlich der Dokumentengröße gibt es kein Einzelprojekt mit hoher Ähnlichkeit. Damit stellt sich für ein ILM-Regelwerk die Frage, ob die Projektzugehörigkeit berücksichtigt werden soll.

Typ/Projekt	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
Stück	748	7	404	68	1	25	9	15	1	27	346
Stück %	42,45	0,4	22,93	3,86	0,06	1,42	0,51	0,85	0,06	1,53	19,64

Typ/Projekt	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc
Stück	23	3	19	53	2	2	5	1	1	2
Stück %	1,31	0,17	1,08	3,01	0,11	0,11	0,28	0,06	0,06	0,11

Tabelle 14: Alle Projekte – Dokumentenanzahl nach Dateityp

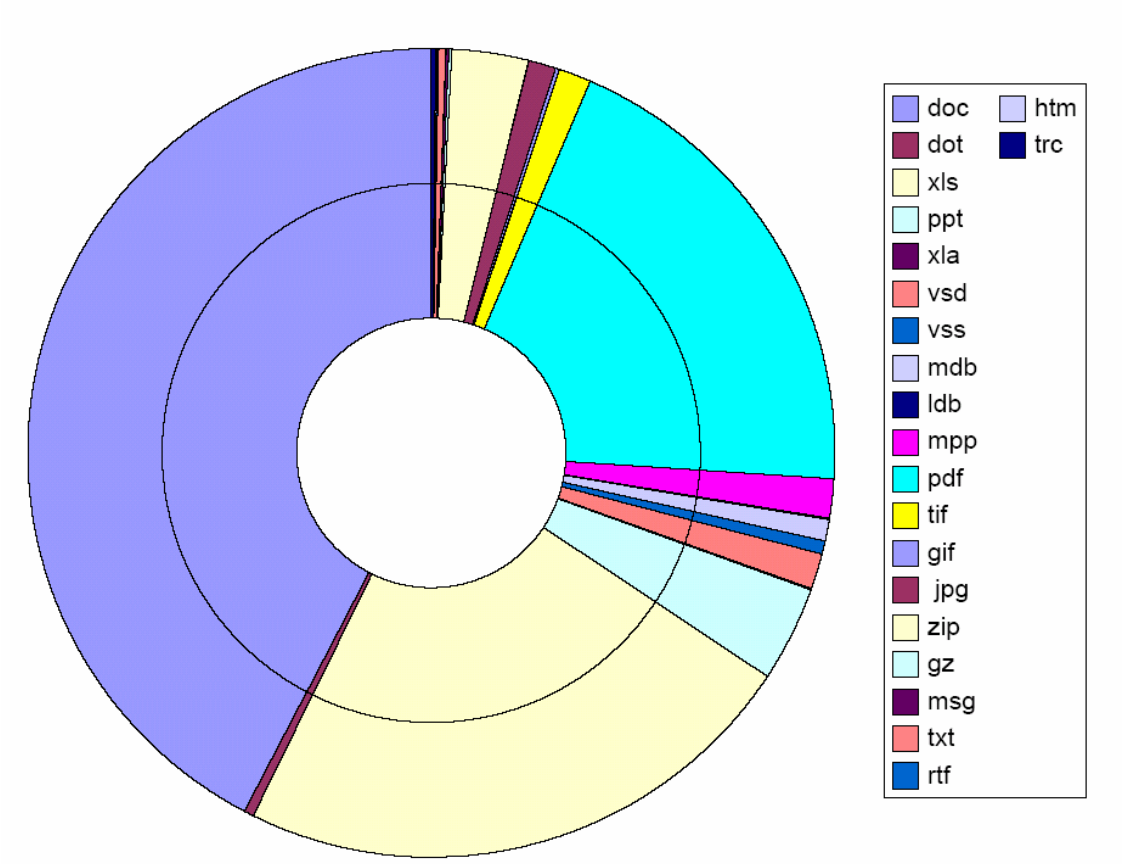

Abbildung 12: Alle Projekte - Dokumentenanzahl nach Dateityp

Typ/Projekt	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
KB	456956	350	95983	91353	35	8294	261	4032	2	6587	82857
KB %	48,51	0,04	10,19	9,7	0	0,88	0,03	0,43	0	0,7	8,8

Typ/Projekt	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc
KB	5293	11	13506	173026	918	71	15	34	8	2480
KB %	0,56	0	1,43	18,37	0,1	0,01	0	0	0	0,26

Tabelle 15: Alle Projekte – Dokumentengröße nach Dateityp

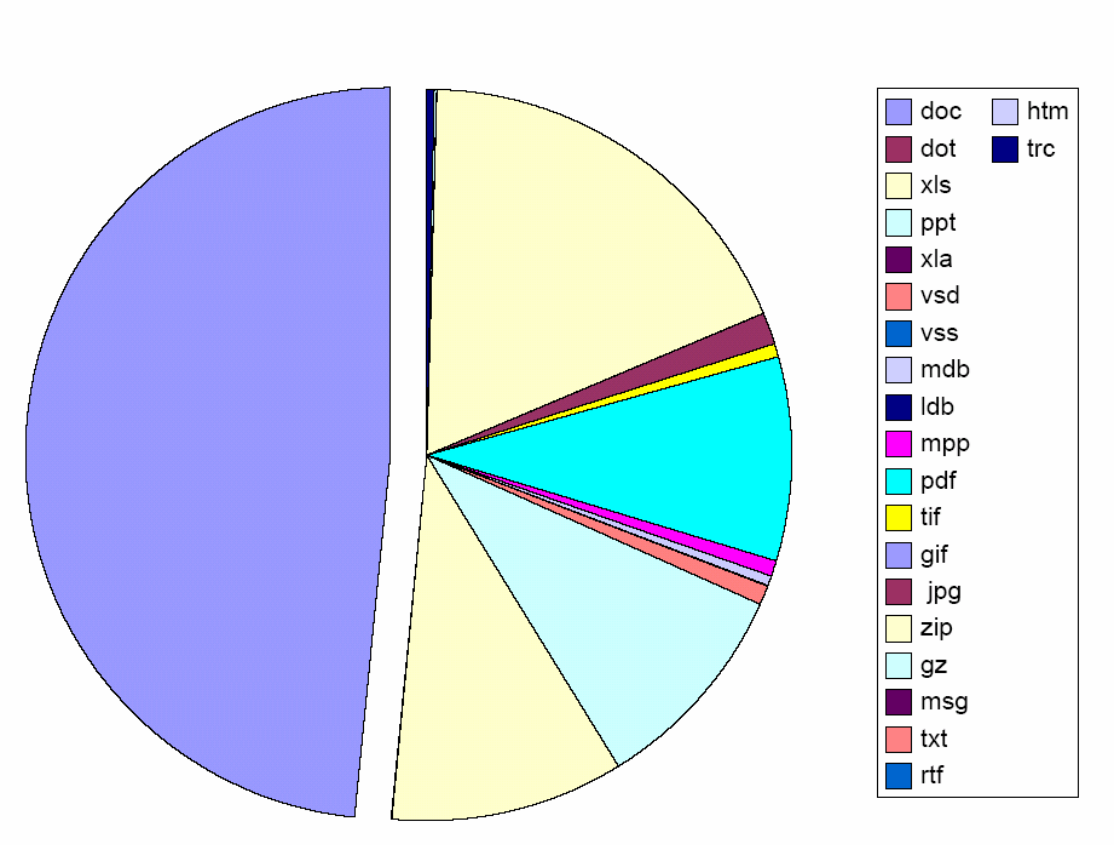

Abbildung 13: Alle Projekte – Dokumentengröße nach Dateityp

Die Inspektion einfacher empirischer Maßzahlen der Daten (Tabellen 16-19)[7] deutet darauf hin, dass die Aufteilung nach Dokumentkategorien stark vom Projekt abhängt. Dokumentenzahl und Speicherbedarf streuen ebenfalls stark zwischen den Projekten. Das könnte darauf hinweisen, dass ein einheitliches Regelwerk ein deutlich geringeres Potenzial hebt[8].

Kategorie	Mittelwert (MW)	Std.abw.	Std.abw./MW (%)
Dokumentenaufbereitung	94,44	83,93	88,87
Dokumentenhaltung	49,67	56,25	113,25
Archivierung	44,56	52,78	118,45

Tabelle 16: Streuung der Dokumentenanzahl

Kategorie	Mittelwert (MW)	Std.abw.	Std.abw./MW (%)
Dokumentenaufbereitung	61873,78	96632,72	156,18
Dokumentenhaltung	11844,89	13346,55	112,68
Archivierung	28533,44	37851,5	132,66

Tabelle 17: Streuung der Dokumentengröße

[7]Für die auf Dokumentkategorien entfallenden Anteile wurden gerundete Werte verwendet, was bei verschiedenen Einträgen zu weiteren Rundungsfehlern führte.

[8]Valide Aussagen lassen sich nur mit Verteilungstests ermitteln was den Rahmen dieser Arbeit übersteigt.

Kategorie	Mittelwert (MW)	Std.abw.	Std.abw./MW (%)
Dokumentenaufbereitung	0,59	0,2	33,48
Dokumentenhaltung	0,22	0,11	49,01
Archivierung	0,19	0,12	66,29

Tabelle 18: Streuung der Kategorienanteile bzgl. Dokumentenanzahl

Kategorie	Mittelwert (MW)	Std.abw.	Std.abw./MW (%)
Dokumentenaufbereitung	0,55	0,27	50,22
Dokumentenhaltung	0,16	0,12	72,89
Archivierung	0,29	0,23	78,29

Tabelle 19: Streuung der Kategorienanteile bzgl. Dokumentengröße

Die aggregierte Zugriffsstatistik (Tabelle 20) zeigt, dass mit einem einfachen Regelwerk praktisch kein Potenzial gehoben werden kann. Eine Ausnahme bilden ausgewählte Dateitypen wie LDB, HTM, GIF, TIF und in Grenzen auch die restlichen Dokumentenarten der Kategorie Sonstiges. Zwar verzeichnen diese teilweise Zugriffe in vielen Intervallen, aber die absolute Zahl der Dokumente ist gering. Außerdem handelt es sich bei den Typen häufig um Hilfsdateien mit vornehmlich technischer Bedeutung für die assoziierten Applikationen (zum Beispiel TRC, LDB, siehe Anhang B).

	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
(1, 3)	100	100	100	100	100	100	100	100	100	100	100
[3, 7)	0,9	0	0,92	0	0	0	0	0	0	0	0,89
[3 , 7)	0,88	0	0,84	0	0	2,5	0	0	0	3,33	1,11
[7, 15)	7,86	14,29	9,46	8,33	0	4,17	7,35	1,19	0	28,33	11,68
[15, 30)	8,31	0	9,8	7,09	50	18,75	12,5	0	0	43,33	7,35
[30, 60)	2,03	0	10,98	12,09	0	1,04	0	0,13	0	13,33	2,57
[60, 90)	15,71	28,57	24,6	25,12	50	19,79	25	25,25	0	37,08	15,87
[90, ∞)	10,58	14,29	10,34	16,64	0	15,42	18,75	12,5	0	25,83	7,39

	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc	
(1, 3)	100	100	100	100	100	100	100	100	100	0	
[3, 7)	0	0	0	0,55	0	0	0	0	0	0	
[3 , 7)	0	0	0	0	0	0	0	0	0	0	
[7, 15)	0	0	0	0,55	0	0	0	0	0	0	
[15, 30)	0	0	40	17,03	100	150	0	100	0	100	
[30, 60)	50	0	1,54	2,89	0	0	0	0	0	100	
[60, 90)	0	0	40	21,43	0	200	37,5	100	0	100	
[90, ∞)	0	0	20	13,19	100	100	37,5	0	0	100	

Tabelle 20: Alle Projekte – Relative Zugriffshäufigkeiten nach Dokumenttyp und Zeitintervall

Typ	doc	dot	xls	ppt	xla	vsd	vss	mdb	ldb	mpp	pdf
KB %	48,51	0,04	10,19	9,7	0	0,88	0,03	0,43	0	0,7	8,8
[90, ∞)	10,58	14,29	10,34	16,64	0	15,42	18,75	12,5	0	25,83	7,39
-[90, ∞)	89,42	85,71	89,66	83,36	100	84,58	81,25	87,5	100	74,17	92,61
Potenzial	43,38	0,03	9,13	8,08	0	0,74	0,02	0,37	0	0,52	8,14

Typ	tif	gif	jpg	zip	gz	msg	txt	rtf	htm	trc	Summe
KB %	0,56	0	1,43	18,37	0,1	0,01	0	0	0	0,26	100
[90, ∞)	0	0	20	13,19	100	100	37,5	0	0	100	0
-[90, ∞)	100	100	80	86,81	0	0	62,5	100	100	0	0
Potenzial	0,56	0	1,15	15,94	0	0	0	0	0	0	88,09

Tabelle 21: Potenzial nach Dokumenttyp und gesamt

Ausgehend von den Dokumenten, die nach 90 Tagen keinen Zugriff zu verzeichnen haben, wird nun bestimmt, wie hoch der von ihnen belegte Anteil am Gesamtspeicherbedarf der Stichprobe ist. Tabelle 21 zeigt, dass es ein Potenzial von 88 Prozent des belegten Speichers zu heben gibt. Dieses Ergebnis bestätigt die Untersuchungen von Gibson et al. Aus dem Jahre 1998.

Um die Charakteristika der Zugriffe noch genauer zu ermitteln, wird nachfolgend das Intervall [90,∞) weiter unterteilt.

7 Auswertung aller Projekte auf 400-Tage-Basis

In diesem Kapitel werden zuerst die Zugriffe bis zu 400 Tage nach Erstellung untersucht. Tabelle 22 zeigt, dass ca. 9 Prozent aller Zugriffe zwischen dem 250. und 300. Tag nach Erstellung erfolgt. Eine Erklärung dafür ist nicht eindeutig zu geben. Es kann an unternehmenseigenen Prozeduren der Revision oder des Controllings liegen.

Tage nach Erstellung	Häufigkeit
(1,3)	3574
[3,7)	89
[7,15)	229
[15,30)	255
[30,60)	296
[60,90)	677
[90,120)	63
[120,150)	162
[150,200)	46
[200,250)	146
[250,300)	552
[300,350)	3
[350,400)	11
[400,)	14
Summe	6117

Tabelle 22: Absolute Zugriffshäufigkeiten auf 400-Tage-Basis

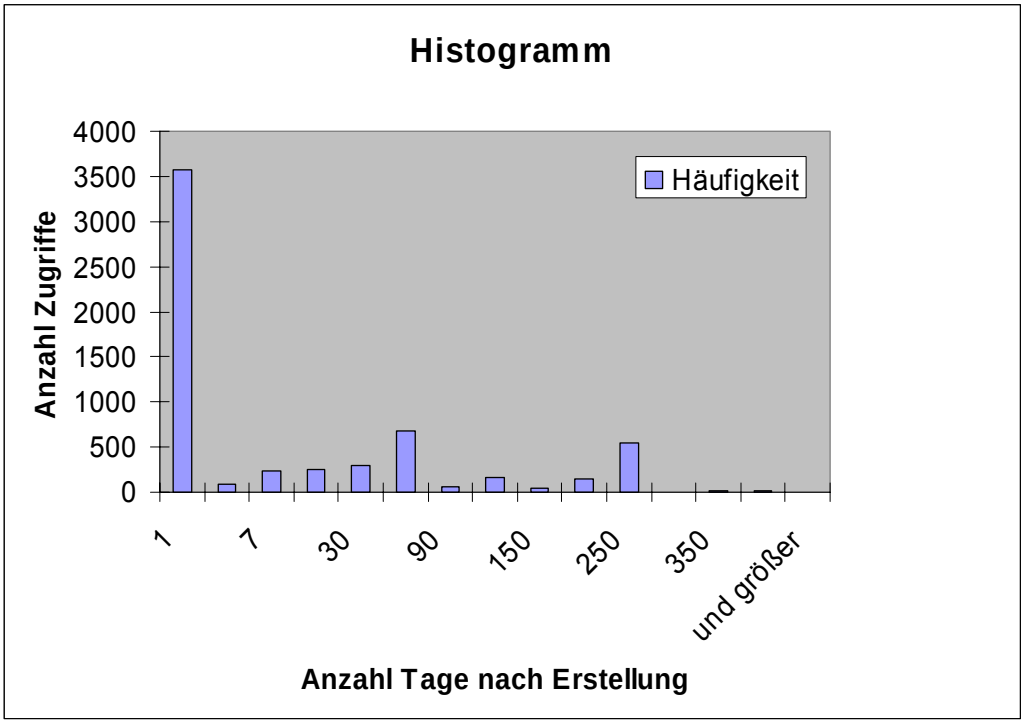

Abbildung 14: Histogramm der absoluten Zugriffshäufigkeiten

Betrachtet man die vergangene Zeit zwischen den Zugriffen, so erkennt man hier eine Häufung im Bereich 200. bis 250. Tag.

Tage nach letztem Zugriff	Häufigkeit
(1,3)	4216
[3,7)	109
[7,15)	179
[15,30)	396
[30,60)	94
[60,90)	613
[90,120)	31
[120,150)	118
[150,200)	26
[200,250)	328
[250,300)	1
[300,350)	6
[350,400)	0
[400,)	0
Summe	6117

Tabelle 23: Absolute Zugriffshäufigkeiten seit letztem Zugriff

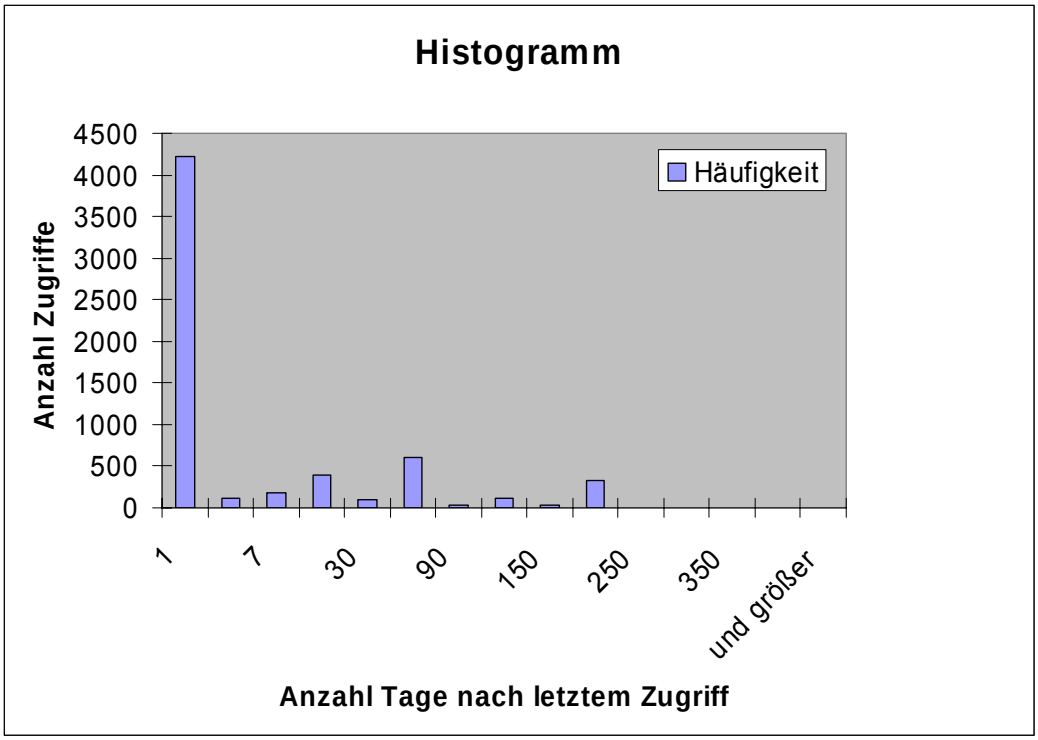

Abbildung 15: Histogramm der absoluten Zugriffe seit letztem Zugriff

Die Betrachtung auf 400-Tage-Basis zeigt, dass, obwohl 89% der Dateien keine Zugriffe 90 Tage nach Erstellung erfahren, man dennoch die Zugriffe über einen längeren Zeitraum als 90 Tage betrachten muss.

8 Zusammenfassung

Im Rahmen des Technical Reports wurde ein Datenbestand eines Unternehmens analysiert, um zu prüfen, ob genügend Potenzial vorhanden ist, so dass ILM nutzbringend angewendet werden kann.

Es wurde ein Potenzial von 90 Prozent identifiziert, welches 90 Tage nach Erzeugung brach liegt. Eine kanonische Regel für die Datenbank lautet wie folgt:

Lösche oder Verdänge alle Dokumente 90 Tage nach Erstellung.

Eine derartige Regel würde allerdings eine Fehlerrate von circa 10 Prozent verursachen, was unvertretbar ist. Die Fehler ziehen sich durch fast alle Applikationen hindurch.

Es zeigt sich, dass die Horison-Studie [5] im Spezialfall der untersuchten Datenbank des DAX-30-Unternehmens nicht aussagekräftig ist.

Insbesondere die betrachteten Intervalle bei Horison sind zu grob. Man muss das Intervall [90, ∞) in weitere Teilintervalle unterteilen und untersuchen werden, wie die Betrachtung auf 400-Tage-Basis zeigt.

Obwohl 89% der Dateien keine Zugriffe 90 Tage nach Erstellung erfahren, finden über 16% der Zugriffe (997 von 6117) nach 90 Tagen nach Erstellung statt. Diese Erkenntnis sollte bei einer ILM-Regelerstellung berücksichtigt werden.

Nur mittels der Unterteilung auf 400-Tage-Basis lässt sich das identifizierte Potential, zumindest teilweise erschließen.

9 Literatur

[1] Michael Peterson. Information Lifecycle Management. A Vision for the Future. http://www.snia.org/tech_activities/dmf/docs/, Juli 2004.

[2] Michael Peterson. ILM Definition and Scope -An ILM Framework. http://www.snia.org/dmf, Juli 2004.

[3] Michael Peterson, Edgar St. Pierre. Information Lifecycle Management Roadmap. http://www.snia.org/tech_activities/dmf/docs/, Oktober 2004.

[4] L. Turczyk. ILM: Vom Schlagwort zur Notwendigkeit. Infor*mation Wissenschaft & Praxis,* September 2004.

[5] Fred Moore. Information Lifecycle Management. http://www.storagetek.com/solutions/technical_papers.html, Juli 2004

[6] M. Satyanarayanan, A Study of File Sizes and Functional Lifetimes, Proceedings of the 8th ACM Symposium on Operating Systems Principles, 1981

[7] Stephan Strange, Analysis of Long-term Unix File Access Patterns for Application to Automatic File Migration Strategies, University of Berkeley, 1992

[8] T. Gibson, E.L. Miller, D.D.E. Long, Long-term File Activity and Inter-Reference Patterns, 24th International Conference on Technology Management and Performance Evaluation of Enterprise-Wide Information Systems, Computer Measurement Group, Anaheim, California, USA, 1998

[9] A. Iamnitchi, M. Ripeanu, Myth and Reality: Usage Behavior in a Large Data-Intensive Physics Projekt, University of Chicago, Chicago, USA, 2002

[10] C. Roadknight, I. Marshall, D. Vearer, File Popularity Characterisation, BT Research Laboratories, Suffolk, UK, 1999