



CRIA - Archivierung von Textinformationen

kippdata informationstechnologie GmbH, Juni 2008

1 Einleitung

Das vorliegende kurze Dokument dient der Darstellung einiger Eigenschaften von CRIA, die für den Erfolg einer unternehmensweiten Archivierungs- und Informationsmanagement-Strategie von Bedeutung sind.

Laut aktuellen Studien liegen etwa 70% aller relevanten Informationen in einem Unternehmen textbasiert und semistrukturiert vor. Viele dieser Informationen werden nicht zentral erfasst und erschlossen. Dies gilt insbesondere auch für den großen Komplex der elektronischen Mail, die personenorientiert bearbeitet und abgelegt wird.

CRIA ist bestens geeignet, gerade auch in großen und heterogenen Umgebungen die Erschließung und Nutzung sämtlicher textbasierter Daten zu ermöglichen. CRIA ist modular aufgebaut und läßt sich problemlos in eine bestehende IT-Infrastruktur eingliedern.

2 Anforderungen an Archivierung

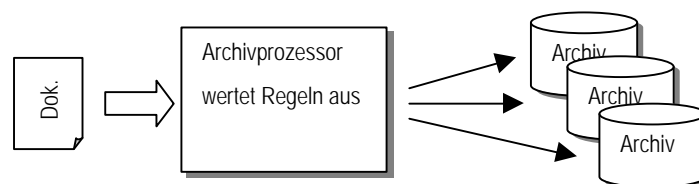
Die Archivierung textbasierter Daten (oder Informationen) stellt den Schlußpunkt der Prozessierung dar. Durch die Archivierung soll sichergestellt werden, daß geschäftsrelevante Informationen auch zu einem späteren Zeitpunkt zugreifbar sind.

Die Motivation hierfür stammt im wesentlichen aus zwei Beweggründen:

1. **Rechtliche Anforderungen:** Eine Vielzahl von Gesetzen und Verordnungen schreibt den Umgang mit geschäftsrelevanten Dokumenten vor. Diese Vorschriften gelten unabhängig davon, in welcher Form ein Dokument vorliegt.
2. **Informationsmanagement:** Die vorhandenen Informationen sollen innerhalb des Unternehmens aus diversen Gründen (Vertrieb, Marketing, Produktentwicklung und andere) nutzbar sein.

Um diese fachlichen Anforderungen erfüllen zu können, muß die Aufbewahrung der Daten gewisse Kriterien erfüllen.

- (1) Die Archivierung muß organisatorische Regeln abbilden. Das bedeutet, daß sie nicht in einem einzigen physischen Archiv erfolgen kann, sondern eine Aufteilung der Inhalte auf unterschiedliche Speicherorte gemäß vorliegender Regeln stattfinden muß.



Die vorstehende Graphik zeigt, wie ein zu archivierendes Dokument von einem Archivprozessor gemäß dort hinterlegter fachlicher (organisatorischer) Regeln in zu einem physischen Data Store geroutet wird.

Beispiele für solche fachlichen Regeln sind das Routing nach Vorgängen oder die Einhaltung von Sicherheitsvorschriften innerhalb einer Organisation.

- (2) Die archivierten Daten müssen jederzeit leicht zugänglich sein, damit sie genutzt werden können. Die Recherche auf den Daten sollte ohne Spezialkenntnisse über die Art der Ablage oder die interne Ablageorganisation erfolgen können.

Die Erschließung der Daten darf nicht statisch sein, da sich Strukturen oder Organisationskriterien im Laufe der Zeit ändern können.

- (3) Die Zusammenstellung von relevanten Daten zu einem oder mehreren Geschäftsvorfällen muß innerhalb der Fristen erfolgen können, die beispielsweise im Rahmen des E-Discovery vor US-amerikanischen Gerichten gefordert werden.

Diese Aufgabe erfordert nicht nur die technische Möglichkeit zur Zusammenstellung von Daten aus unterschiedlichen Quellen, sondern auch die Festlegung der Kriterien, nach denen die Relevanz von Informationen beurteilt wird. Nur so kann der Wert der bereitgestellten Informationen von dritter Seite aus beurteilt werden.

Aus der obigen kurzen Liste folgt schon, daß die Archivierung von textbasierten Informationen nicht ohne deren inhaltliche Erschließung erfolgen kann. Das ist keine Überraschung, denn auch klassische Aktenarchive bestehen nicht aus der unterschiedslosen Aufbewahrung der Rohdaten, sondern werden paginiert und durch Hinzufügen von Schlüsselbegriffen erschlossen. Zu einem Aktenarchiv gehört immer auch ein Katalog oder Index, der das Auffinden von Information ermöglicht.

Elektronische Dokumente werden typischerweise ebenfalls manuell erfasst, wenn sie in ein Dokumenten-Management-System (DMS) eingestellt werden. Der erfassende Benutzer muß eine Maske ausfüllen, in der diverse Attribute abgefragt werden. Diese werden zum Aufbau eines Index genutzt, auf dem spätere Recherchen abgearbeitet werden.

Es ist seit langem durch Studien belegt, daß die manuelle Klassifizierung von Texten – auch durch Spezialisten – nicht gut in dem Sinne ist, daß sie unabhängig von der klassifizierenden Person zu den gleichen Ergebnissen führt: Wenn zwei Personen den gleichen Textkorpus klassifizieren, so führt das in der Regel nur zu Übereinstimmungen in der Größenordnung von etwa 30%. Die Klassifizierung von elektronischen Texten für ein DMS erfolgt aber in der Regel nicht durch Personen, die für diese Aufgabe speziell ausgebildet sind. Es handelt sich um eine Aufgabe, die zusätzlich zu den eigentlichen Tätigkeiten erfüllt werden muß. Da zudem der unmittelbare Nutzen der Klassifizierung nicht offenbar ist, wird die Qualität des Ergebnisses schwankend und unzuverlässig sein.¹

Neben diesen qualitativen Aspekten sind auch quantitative Hindernisse gegen die manuelle Klassifizierung zu beachten: Die große Menge an Rohdaten, die aus Mail, internen Texten und gescannten Dokumenten besteht, welche alle nach gleichen oder vergleichbaren Regeln klassifiziert werden müssen, macht die manuelle Durchführung praktisch unmöglich.

Erforderlich ist die Implementierung eines Verfahrens, das unbedingt die folgenden fachlichen Kriterien erfüllt:

- (1) Die Erfassung aller textbasierten Informationen unabhängig von der Quelle muß sichergestellt sein.
- (2) Die Klassifizierung der Rohdaten muß nach einheitlichen Regeln erfolgen.
- (3) Die Benutzerschnittstelle für die Recherche muß ohne Spezialwissen über die Klassifizierung oder den erstellten Index nutzbar sein.
- (4) Die Archivierung der klassifizierten Daten muß regelbasiert auf unterschiedlichen Medien erfolgen können.
- (5) Die spätere Zusammenstellung von Informationen aufgrund ad hoc definierter Kriterien muß vom Benutzer ohne Aufwand initiiert werden können.

Von großer Bedeutung ist, daß die Etablierung einer Archivierungslösung nicht ausschließlich aus der Einführung einer Software besteht, sondern in großem Maße die Feststellung der internen Informationsflüsse, der unterstützenden weitergehenden Informationen sowie der typischen Nutzungsmuster erfordert.

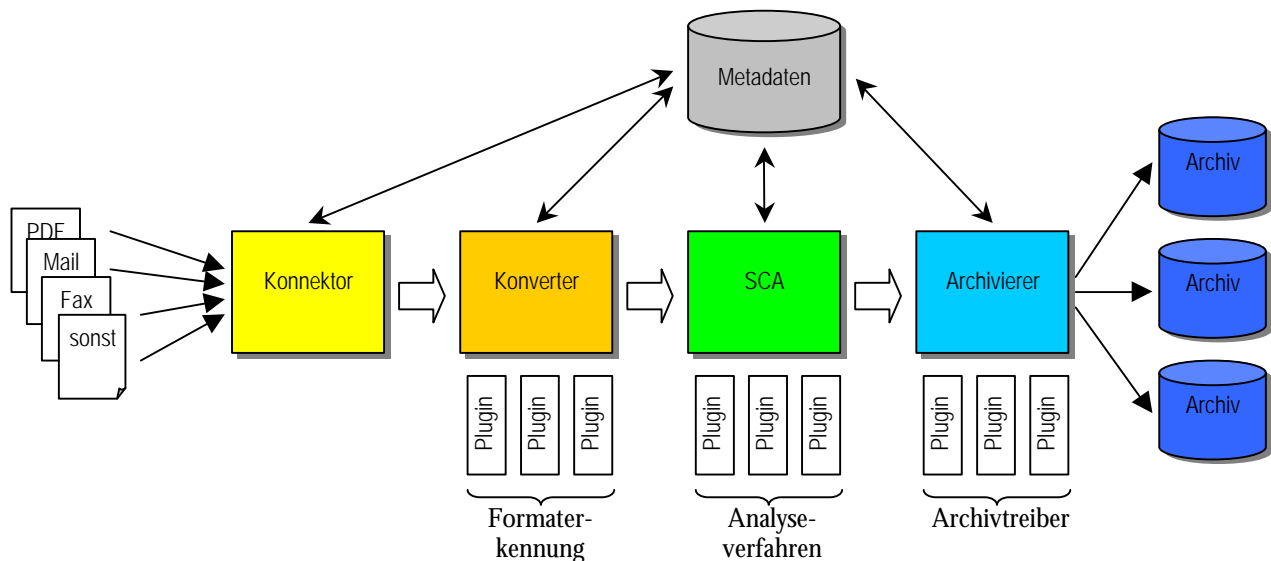
3 Eigenschaften von CRIA

Das Produkt CRIA (Common Record Information Application) stellt eine moderne Plattform für die Erfassung, das Routing und die Archivierung von (textbasierten) Daten dar.

CRIA ist modular aufgebaut und skaliert optimal in einer modernen serviceorientierten IT-Landschaft. Alle Komponenten von CRIA können einzeln oder in Gruppen auf die unterliegenden Systeme ausge-

¹ Natürlich kann man argumentieren, daß jeder Erfasser später einmal mit einer gewissen Wahrscheinlichkeit Nutznießer einer guten Erfassung wird, wenn er selber eine Recherche auf den gesammelten Daten vornimmt. Studien zeigen jedoch, daß dieser ferne Nutzen, der auch nur potentiell ist, gering gewertet wird im Vergleich zu der aktuell anliegenden zusätzlichen Arbeit.

bracht werden. Performanceanforderungen werden durch die Vervielfachung einzelner Komponenten erfüllt.



Die vorstehende Abbildung zeigt die einzelnen Komponenten von CRIA und ihr Zusammenspiel bei der Erfassung, Klassifizierung und Archivierung von textbasierten Daten.

Die farblich gekennzeichneten rechteckigen Komponenten stellen die wesentlichen Bestandteile von CRIA dar. Die drei mit Archiv bezeichneten Objekte rechts außen in der Abbildung sind Data Stores, die vom Archivierer angesprochen werden. Diese Data Stores gehören nicht zu CRIA, sondern sind Bestandteile der IT-Infrastruktur um CRIA herum.

- (1) Der **Konnektor** stellt die Verbindung zwischen einer Datenquelle und CRIA dar. Datenquellen können Mailserver, Fileserver, Faxserver aber auch bereits strukturierte Ablagen wie Datenbanken oder DMS sein. Der Konnektor stellt einige (konfigurierbare) Metadaten bereit.
- (2) Der **Konverter** wandelt die Formate der Daten für die weitere Verarbeitung und stellt gleichzeitig eventuell vorhandene Strukturinformationen als weitere Metadaten bereit.
- (3) Der **SCA** (Scaleable Categorization Analyzer) klassifiziert die Daten. Er verwendet verschiedene Verfahren zur Extraktion von Informationen aus dem vorliegenden Dokument und nutzt hierfür auch externe Informationsquellen innerhalb des Unternehmens.
Der SCA ist durch Plugins erweiterbar und kann so an die besonderen Anforderungen der jeweiligen Situation optimal angepaßt werden.
- (4) Der **Archivierer** ist das Service-Interface zwischen CRIA und den realen Archiven. Seine Archiv-Plugins stellen die Verbindung her und ermöglichen die Übergabe des Archivgutes unabhängig von der jeweiligen technischen Ausprägung des jeweiligen Data Stores.
- (5) Das **Metadaten-Repository** sammelt alle Metadaten der erfassten und klassifizierten Daten. Es stellt die Services für die Recherche, die Einbindung von Workflow-Systemen und das Reporting bereit.

CRIA erfüllt alle Anforderungen an ein modernes Informationsmanagement-System für textbasierte Daten. Die variable Ausgestaltung des SCA gestattet die Extraktion von Informationen aus Dokumenten unterschiedlicher Art und Herkunft unter Verwendung aller in einer Organisation vorhandenen und bereitstellbaren Zusatzdaten. Die Auskopplung der physischen Archive und Bereitstellung entsprechender Service-Schnittstellen adressiert zum einen die Forderung nach regelgesteuerter Ablage von Dokumenten und stellt einen maximalen Schutz für bereits getätigte Investitionen im Bereich der Datenspeicherung und der Dokumentenablage dar. Die Flexibilität der Lösung ermöglicht die Kommunikation mit anderen Systemen im Rahmen serviceorientierter Architekturen und erlaubt insbesondere die Kopplung mit DMS-Installationen.