



CRIA im Überblick

kippdata informationstechnologie GmbH, Juni 2008

1 Einleitung

E-Mail ist heute ein zentrales und unverzichtbares Kommunikationsmittel in Unternehmen und Verwaltungen jeder Größe. Rechtlich und organisatorisch relevante Informationen werden über dieses Medium ausgetauscht. Hierdurch entsteht zunehmend der Bedarf nach einer angemessenen Aufbewahrung, einem umfassenden Reporting für das Management und nach Möglichkeiten der Klassifizierung aller für einen Geschäftsvorfall relevanten E-Mails.

Hierbei sind die Klassifizierung sowie die Verfahren der Aufbewahrung und letztendlichen Vernichtung von E-Mails stark abhängig von den jeweiligen Regeln einer Organisation (Unternehmen oder Verwaltung).

E-Mails nehmen in einer Organisation unterschiedliche Aufgaben wahr. Sie dienen der internen Steuerung von Abläufen ebenso wie der Kommunikation mit Kunden, Partnern und Lieferanten. Für alle diese Anwendungsfälle ist es erforderlich, daß entsprechende Verfahren zur Behandlung der enthaltenen Informationen angewandt werden.

Unsere Lösung – die **Common Record Information Application**, kurz CRIA – setzt neue Maßstäbe im Umgang mit E-Mails, sie ist ein E-Mail-Management der nächsten Generation, indem sie auf bewährten Verfahren aufsetzt und diese um neuartige Ansätze ergänzt. So entsteht ein System, das höchsten Ansprüchen gerecht wird.

2 Die fachliche Lösung

2.1 Anforderungen

Die zentrale Anforderung an Mail-Management besteht in der Auflösung der Bindung zwischen einer Mail und ihrem Empfänger. E-Mails und ihre Anhänge müssen jederzeit von jeder berechtigten Person eingesehen werden können. Die Zugreifbarkeit darf nicht durch die individuellen Ablageverfahren einzelner Anwender behindert oder beeinträchtigt werden.

E-Mails müssen automatisch nach inhaltlichen Kriterien in ein Kategorisierungsschema verbracht werden. Eine solche Einordnung muß die Zuweisung einer Mail auch zu mehreren Kategorien gleichzeitig zulassen. Das Kategorisierungsschema muß unter Berücksichtigung der organisationsinternen Verfahren erstellt und angepaßt werden können. Typische Kategorisierungsschemata sind die Klassifizierung nach

- Kunden und Lieferanten
- interner Organisation: Einkauf, Finanzen, Entwicklung, Vertrieb, ...
- Projekten und Mandaten

wobei durchaus auch mehrere Schemata gleichzeitig sinnvoll sein können.

Die Ablage der klassifizierten E-Mails und ihrer Anhänge muß so erfolgen, daß organisatorische Ablage-regeln hinsichtlich der Verteilung von Daten auf mehrere Archive in der Lösung realisiert werden können. Die Regeln müssen auch die Klassifizierung der Mails berücksichtigen.

Das Löschen von archivierten Informationen muß gemäß den hinterlegten Regeln automatisch möglich sein. Um den vielfältigen regulatorischen Vorgaben zu genügen, muß jeder Löschvorgang dokumentiert werden. Die Löschprotokolle müssen für spätere Zwecke zugänglich sein.

2.2 Realisierung

Die Realisierung durch CRIA erledigt alle in 2.1 genannten Anforderungen. Die folgende Abbildung skizziert grob die Abläufe von CRIA zur Kategorisierung und Ablage von E-Mails.

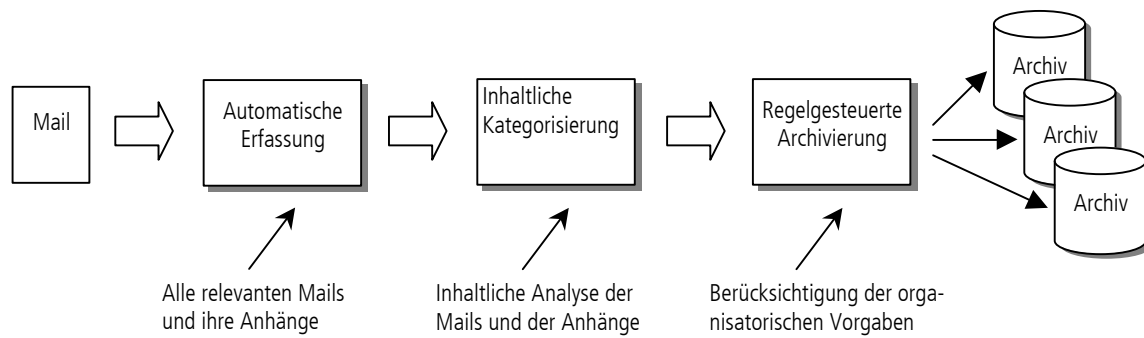


Abbildung 1: E-Mail-Management mit CRIA im Überblick

Alle E-Mails werden zusammen mit ihren Anhängen ohne manuelles Zutun der Benutzer erfasst. Unabhängig von der Art der Anhänge werden diese in der weiteren Verarbeitung mit berücksichtigt. Die erfassten Mails werden durch die Kategorisierung von CRIA inhaltlich klassifiziert und einer oder mehreren Kategorien zugewiesen. Diese Kategorien spiegeln die kundenspezifische Organisation von Dokumenten wider (vgl. die Aufzählung in 2.1).

Jede E-Mail und jeder Anhang wird aufgrund der von CRIA erfassten Informationen und unter Berücksichtigung von kundenspezifischen Ablageregeln in einem Archiv abgelegt. CRIA ermöglicht das einfache und sichere Wiederauffinden aller abgelegten Daten zu einem späteren Zeitpunkt durch entsprechend autorisiertes Personal. Ein ausgefeiltes regelbasiertes Berechtigungskonzept schützt sensible Daten vor unberechtigtem Zugriff.

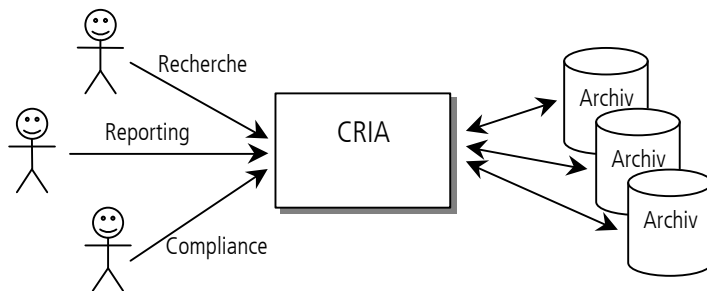


Abbildung 2: Interaktion der Anwender mit CRIA

Die vorstehende Abbildung zeigt die Benutzung von CRIA durch unterschiedliche Benutzer. Die Recherche von Informationen zur Abklärung von Vorgängen in der Vergangenheit, das Reporting von Vorgängen und die Sicherstellung von Compliance-Auflagen erfolgt jeweils durch entsprechend berechtigtes Personal und gewährleistet so die gleichzeitige Berücksichtigung betrieblicher Belange sowie des Datenschutzes und anderer Schutzvorschriften.

Aufgrund der vorgangsbezogenen Verarbeitung und Ablage ist sichergestellt, daß alle Informationen zur Verfügung stehen.

3 Die technische Realisierung

Im folgenden wird die technische Realisierung von CRIA erläutert. Die Lösung ist modular aufgebaut und paßt sich optimal in moderne serviceorientierte IT-Landschaften ein. Die nachfolgende Abbildung zeigt im Überblick den Aufbau der verschiedenen CRIA-Komponenten, die durch ihr Zusammenspiel ein effizientes E-Mail-Management gewährleisten:

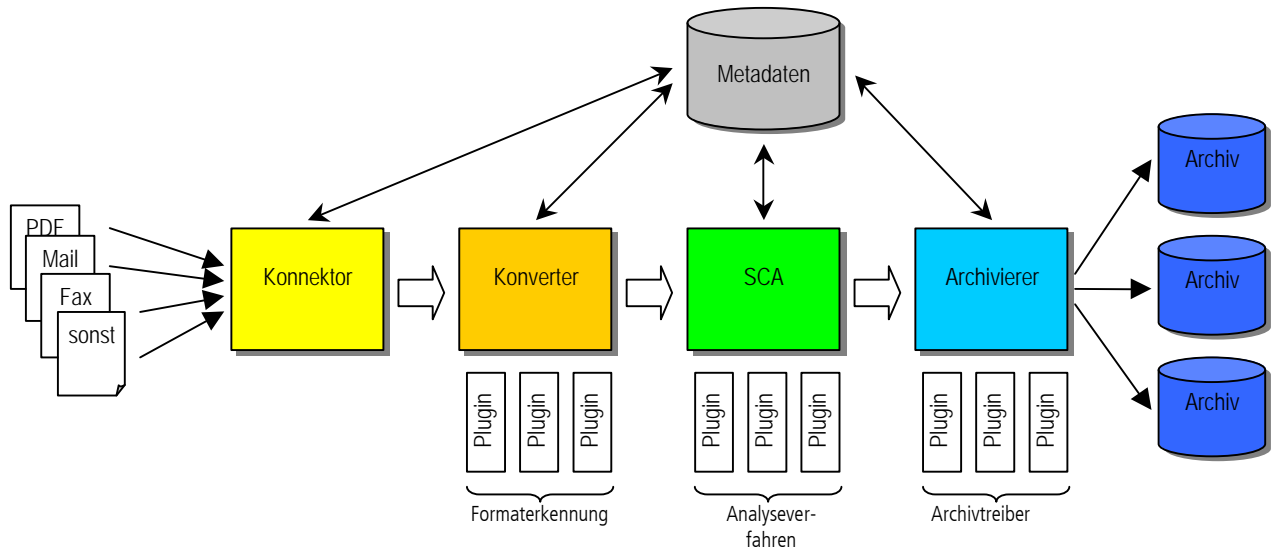


Abbildung 3: Der Aufbau von CRIA

Die einzelnen Komponenten und ihr Beitrag zur Gesamtlösung werden in den Unterkapiteln 3.1 bis 3.5 einzeln besprochen.

Die vorgangsbezogene Sicht auf E-Mails ist der zentrale Angelpunkt der gesamten Lösung. Zu diesem Zweck zweigt CRIA alle Mails zusammen mit ihren Anhängen direkt am Mailserver ab und versieht sie automatisch mit den Metadaten, die für die Zuordnung zu einem Geschäftsvorfall erforderlich sind. Die bestehende Mail-Infrastruktur muss dabei nur minimal angepasst werden, was einen großen Vorteil bei der Implementierung von CRIA darstellt: Die Benutzer können ihre vorhandenen Mailprogramme weiter benutzen.

Im Zuge der Abzweigung von E-Mails am Mailserver wird von jeder E-Mail (interne wie externe E-Mails) eine Kopie in CRIA erzeugt, auf die das System später für Reports und Recherchen zurückgreifen kann. Diese Kopien werden außerdem noch so zerlegt, daß die häufig anzutreffenden Anhänge (Attachments) abgespalten und separat aufbewahrt werden. Damit kann zum einen erreicht werden, daß Anhänge, die an einen großen Verteiler geschickt worden sind, nur einmal aufbewahrt werden, und zum anderen, daß auch diese Anhänge einer weiteren Verarbeitung zugänglich gemacht werden. So erübrigt sich die manuelle Ablage solcher Anhänge auf der Festplatte, weil sie mittels CRIA besser gefunden werden können.

3.1 Konnektor

Der Konnektor ist diejenige Komponente von CRIA, die alle zu verarbeitenden E-Mails vom Mailserver aufnimmt und an das Kernsystem weiterleitet. Es handelt sich um eine leichtgewichtige Komponente, die für jeden Typ von Mailserver (Exchange, Notes/Domino, GroupWise und andere) die spezifische Anbindung an CRIA vornimmt.

Der Konnektor nimmt hierbei nicht nur die Mails samt Anhängen entgegen, sondern versieht sie mit ersten Metadaten wie zum Beispiel einem Zeitstempel. Damit kann später der Zeitpunkt des Mailempfangs rekonstruiert und nachgewiesen werden. Das ist für forensische Fragen im Rahmen von Rechtsstreitigkeiten von erheblichem Belang.

3.2 Konverter

E-Mails und ihre Anhänge liegen in einer Vielzahl von verschiedenen Datenformaten vor. Darunter befinden sich unter anderem HTML, Microsoft Word, PDF oder TIFF. Diese Formate müssen in ein ge-

meinsames Format umgewandelt werden, das für die Kategorisierung verwendet werden kann. Da die Anzahl der umzuwandelnden Formate sehr stark von dem jeweiligen Einsatzszenario abhängig ist, ist der Konverter bei Bedarf erweiterbar und stellt über Plugins die jeweils erforderliche Funktionalität zur Verfügung.

Alle oben aufgeführten Datenformate stellen aber nicht nur Textinhalte bereit, sondern können darüberhinaus eine Vielzahl von zusätzlichen Informationen liefern: Ein Word-Dokument beispielsweise enthält Informationen über den Verfasser, das Datum der letzten Änderung und über die innere Struktur des Textes (Kapitelüberschriften, Inhaltsverzeichnisse, Index). Für eine semantische Analyse sind diese Strukturinformationen von großem Nutzen, weswegen der Konverter sie extrahiert und für die spätere Verwendung weiterleitet.

3.3 Scaleable Categorization Analyzer (SCA)

Eine zentrale Eigenschaft von CRIA ist die inhaltliche Analyse von E-Mail und ihren Anhängen. Diese Operation wird auch als Kategorisierung oder Klassifizierung bezeichnet und ist die Grundlage für die automatische Gruppierung von Textdokumenten zu thematischen oder organisatorischen Komplexen.

Die inhaltliche Analyse von Texten geht weit über die übliche Suche nach Schlagworten, die auch als Volltextsuche bekannt ist, hinaus. Volltextsuche ist auch nicht ausreichend für die Zuordnung von Dokumenten zu Themenkomplexen: Einerseits werden gleiche thematische Konzepte von verschiedenen Autoren auf sehr unterschiedliche Weise unter Verwendung unterschiedlicher Begriffe ausgedrückt. Andererseits kann ein Wort je nach Kontext unterschiedliche Konzepte ausdrücken.¹

Die Klassifizierung von Texten wie E-Mail erfordert darüberhinaus Verfahren, die über die Anwendung der klassischen Vektorraum-Algorithmen wie kNN oder SVM hinausgehen. Diese Verfahren bilden die semantische Ähnlichkeit von Texten als Abstand von Vektoren ab und sind damit für gewisse Probleme auch gut geeignet. Speziell E-Mails aber haben einige Besonderheiten, die es zu berücksichtigen gilt:

- Eine einzelne E-Mail ist in der Regel sehr klein.
- Der Text der E-Mail eignet sich nicht gut für die Verwendung klassischer Analyseverfahren².
- Die Anhänge müssen berücksichtigt werden.
- Aber: Anhänge können auch irreführend sein.
- ...

Die (unvollständige) Aufzählung der Schwierigkeiten, die zu überwinden sind, legt einen völlig neuen Ansatz nahe, der von CRIA umgesetzt wird.

Zur Klassifizierung einer E-Mail werden unterschiedliche Verfahren herangezogen, die in sogenannten Analysis Engines (AEs) bereitgestellt werden. Jedes dieser Verfahren ist optimiert für die Auswertung spezieller Informationen, die in einer E-Mail vorhanden sind. Neben der üblichen Berücksichtigung der Adressinformationen aus dem Mail-Header und der semantischen Analyse des Inhaltes können Anrede, Grußformel und die im Geschäftsverkehr erforderlichen formalen Informationen über die Organisation des Absenders ausgewertet werden. Darüberhinaus kann auch das kommunikative Umfeld der zu klassifizierenden Mail Berücksichtigung finden.

Die angewandten Einzelverfahren nutzen je nach Teilaufgabe ganz unterschiedliche Ansätze, die von der komplexen Mustererkennung samt Abgleich mit externen Informationsquellen wie Kundendatenbank oder Corporate Directory bis zu den verschiedenen statistischen Verfahren reichen. Alle Einzelanalysen führen jeweils zu Hinweisen für die Klassifizierung.

Die Hinweise werden gesammelt und zu einer globalen Klassifizierung zusammengefaßt. Auch diese Zusammenfassung der Einzelergebnisse zu einer Gesamtanalyse ist eine Besonderheit von CRIA: Die angewandten Verfahren nutzen die wachsende Kenntnis des Systems über gute Klassifizierungen, um die unterschiedlichen Teilergebnisse zu bewerten und zu gruppieren. Auf diese Weise können typische Probleme anderer Systeme vermieden werden: Beispielsweise kann der Wert der Betreff-Information einer Mail a priori nicht festgestellt werden. Es ist vielmehr erforderlich, daß diese Information kontextabhängig berücksichtigt wird. Dies wird von CRIA geleistet.

¹ Ein Beispiel ist das Wort Bank, das sowohl ein Möbelstück als auch ein Geldinstitut bezeichnen kann.

² Diese sind für die Analyse großer Texte wie beispielsweise Zeitungsartikel optimiert.

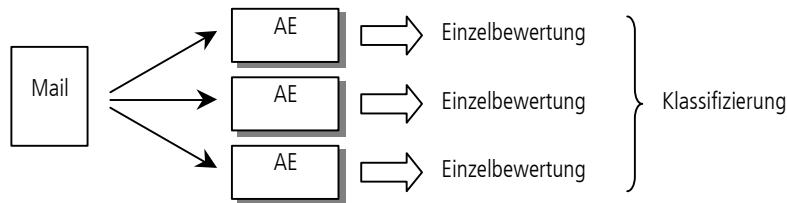


Abbildung 4: Schematische Darstellung der Klassifizierung von E-Mails im SCA

Die vorstehende Abbildung zeigt die Klassifizierung schematisch. Hinter den einzelnen Analysis Engines (AEs) verbergen sich die unterschiedlichen Verfahren; die folgende (auszugsweise) Konfiguration ist typisch für eine Realisierung im Kundenumfeld, wenn eine Klassifizierung nach Kunden und Lieferanten vorgenommen werden soll:

- Identifizierung aller Personennamen und Zuordnung zu Kunden- oder Lieferantendaten gemäß Datenbank
- Identifizierung von bekannten Kunden- oder Lieferantennamen ebenfalls aus der Datenbank
- Identifizierung von Mustern für Angebote, Aufträge sowie Rechnungen und Abgleich mit der Buchhaltung
- Ähnlichkeit einer Mail zu vorhandenen bereits klassifizierten Mails
- Häufigkeit von Begriffen
- ...

Diese Liste läßt sich je nach erforderlicher Präzision und vorhandenen Informationsquellen erweitern. Jede Bewertung liefert einen Beitrag zur Gesamtklassifizierung.

Alle Komponenten können einzeln konfiguriert werden und ermöglichen so die optimale Anpassung an die jeweils vorliegende Aufgabe. Trotzdem ist es bereits mit wenigen Standardeinstellungen möglich, daß gute Ergebnisse erzielt werden.

3.4 Archivierer

Der Archivierung von geschäftsrelevanten E-Mails entsprechend den hinterlegten Regeln kommt eine große Bedeutung zu. Je nach Anforderung kann ein Archiv über eine komplexe logische Struktur verfügen, die auch in der Ablage durch CRIA abgebildet wird.

Der Archivierer nimmt ein zu archivierendes Objekt (eine Mail oder einen Anhang) auf und wertet die vorhandenen Metadaten aus, um die Zuordnung zu einem physischen Speichersystem vornehmen zu können. Somit können die unterschiedlichsten Speichersysteme an CRIA als Archivkomponenten angeschlossen werden. Die folgende Abbildung zeigt das Verfahren schematisch:

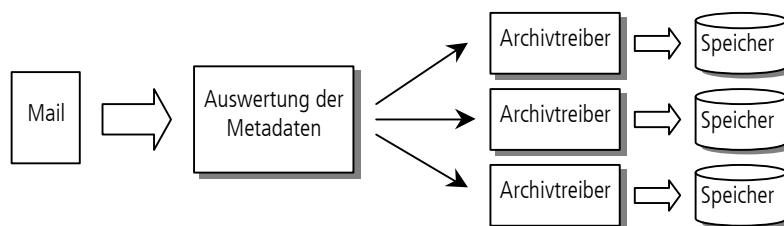


Abbildung 5: Schematische Darstellung der Archivierung

Die Speicherkomponenten ganz rechts in der vorstehenden Abbildung stellen die physischen Komponenten zur Ablage der Informationen dar. Es kann sich um ganz unterschiedliche Realisierungen handeln:

- Relationale Datenbanken (Oracle o.a.)
- HSM-Systeme³

³ HSM steht für Hierarchical Storage Management und bezeichnet stukturierte Speichersysteme typischerweise aus Platten und Bändern zur Aufbewahrung großer Datenmengen. Solche Systeme sind ideal geeignet, um für ein Archiv genutzt zu werden.

- Bandroboter
- Plattensysteme
- Dokumenten-Management-Systeme wie Windream, SAP DMS oder andere
- ...

Aus Sicht von CRIA liegt ein globales Archiv vor, in das die Mails verbracht werden. Je nach Anforderung beim Kunden kann regelbasiert das zu archivierende Objekt beispielsweise auf ein Plattensystem, einen Bandroboter oder auch in ein DMS verbracht werden.

Weiterhin kann sich hinter einem Archivtreiber auch ein Speicher verbergen, der nur bei entsprechender Autorisierung zugänglich gemacht wird. Das läßt sich durch ein Bandgerät realisieren, dessen Bandkassetten jeweils in kurzen Intervallen entfernt und in einen Panzerschrank verbracht werden. Die Anforderung einer derart gespeicherten Mail führt dann zu einer entsprechenden Nachricht an den Benutzer.

Ebenso lassen sich Datenlöschungen als Archive realisieren, die die Daten nicht wirklich aufbewahren, sondern nur ein Löschprotokoll bereitstellen, das später dem Nachweis der ordnungsgemäßen und regelgerechten Vernichtung dient.

Die Nutzung eines DMS als Archiv-Speicher ermöglicht darüberhinaus die interessante Verbindung der Klassifizierung von Mails durch CRIA mit bereits vorhandenen Archivierungsverfahren für Dokumente.

3.5 Weitere Komponenten

Die in den Unterkapiteln 3.1 bis 3.4 vorgestellten Komponenten stellen den technischen Kern der Lösung dar. Für die Benutzung durch Anwender ist aber die Bereitstellung von Benutzerschnittstellen erforderlich. CRIA nutzt für alle Interaktionen der Anwender mit dem System konsequent den Zugang über einen Web-Browser. Dadurch wird erreicht, daß alle erforderlichen Funktionalitäten an jedem Arbeitsplatz zur Verfügung stehen, ohne daß auf den Arbeitsstationen zusätzliche Software installiert werden muß. Dies ist für die Akzeptanz der Lösung bei Benutzern und IT-Verantwortlichen von Bedeutung.

Zur Zeit stellt CRIA folgende Benutzerschnittstellen zur Verfügung:

- Darstellung von Mails entsprechend ihrer Klassifizierung
- Ad hoc-Beauskunftung des Archivs (Suche nach Mails)
- Reporting: Graphische Darstellung von Kennzahlen zur Mail-Kommunikation
- Verwaltung der Zugangsberechtigungen

Diese Benutzerschnittstellen sind einfach zu bedienen und orientieren sich an den gängigen Verfahren. Dadurch werden Anwender schnell mit der Bedienung vertraut und finden sich im System zurecht. Eine Benutzerschulung ist nicht erforderlich.

Die Organisation und Verwaltung von Berechtigungen ist eine wichtige Infrastruktur-Eigenschaft jeder Enterprise-Lösung. Auch CRIA besitzt ein Rechte- und Rollenkonzept, das die Regelung der Zugangs- und Zugriffsberechtigungen auf die gespeicherten Informationen erlaubt:

Abhängig von den Metadaten erlaubt oder verwehrt CRIA einem angemeldeten Benutzer den Zugriff auf die im Archiv abgelegten Informationen. Zur Abbildung der Berechtigungen können Regeln definiert werden. Die folgende Abbildung stellt das Zusammenspiel von Anfragen, Ergebnissen und den Zugriffsregeln dar:

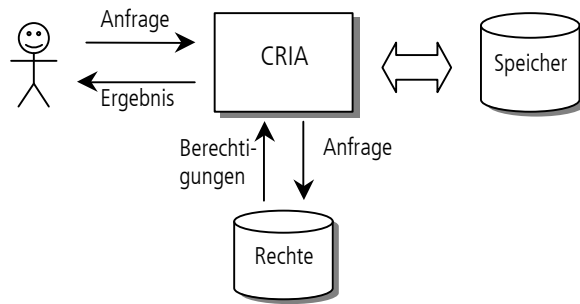


Abbildung 6: Realisierung des Rechte- und Rollenkonzeptes in CRIA

Aus den festgelegten Rechten ergibt sich also, welche Informationen ein angemeldeter Benutzer tatsächlich sehen darf. Das ist insbesondere deswegen von größter Bedeutung, weil Mails vorgangsbezogen verwaltet – und auch angezeigt – werden sollen, darüber aber nicht Vertraulichkeits- oder Datenschutzregelungen verletzt werden dürfen.

Die Ablage der Rechte und Rollen erfolgt, ebenso wie die Ablage der Metadaten, in einer relationalen Datenbank. Dort werden alle Meta-Informationen zu den archivierten Mails abgelegt. Dies hat zur Folge, daß nicht nur die Beauskunftung von Informationen schnell erfolgt, sondern auch bei großen Datenmengen mehrere Instanzen von CRIA gleichzeitig auf dieses Repository zugreifen können.

3.6 Berücksichtigung moderner Software-Entwicklung

CRIA ist konsequent nach den Regeln der modernen Software-Entwicklung konzipiert und realisiert worden. Alle beteiligten Personen verfügen über langjährige Erfahrung im Entwurf und der Realisierung anspruchsvoller Softwaresysteme.

CRIA ist modular aufgebaut und besteht aus Services, die über ein oder mehrere Systeme verteilt werden können. Damit kann die Lösung an unterschiedliche Mailaufkommen von der kleinen Installation mit einhundert Anwendern bis zu großen Konzernen angepaßt werden. Die Komponenten passen sich in eine serviceorientierte Infrastruktur optimal ein.

Alle Komponenten können über Interfaces erweitert werden. Diese Erweiterbarkeit erlaubt sowohl die schnelle Reaktion auf Kundenwünsche (schnelle Bereitstellung neuer Releases) als auch Entwicklung von zusätzlichen Komponenten durch Dritte. Hier können Partner für sich einen Mehrwert erzeugen, der über den Vertrieb von Lizenzen hinausgeht. Aus Marketingsicht führt das zu einer schnelleren Verbreitung bei technikaffinen Partnern und Kunden.