Ausgewählte Applikationen mit Oracle UCM am KIT

Ronny Wörl
KIT
Karlsruhe

Schlüsselworte:

BatchLoader, Berechtigungskonzept, Bilddatenbank, Bildformat, Contentmanagementsystem, Dokumenten-lebenszyklus, Dokumentenmanagementsystem, ImageAlchemy, ImageMagick, Inbound Refinery, KIT, Medienserver, Metadaten, Oracle UCM, SCC

Einleitung

Das Steinbuch Centre for Computing (SCC) bietet innerhalb des Karlsruhe Institute of Technology (KIT) verschiedene Dienste an. Unteranderem die mit Oracle UCM entwickelten Dokumentenmanagementsysteme und verschiedene Webportale.
Dieser Vortrag beschäftigt sich mit dem Einsatz von Oracle UCM innerhalb des KITs, den Hintergründen und Erfahrungen.
Zunächst werden einige Begriffe rund um das Thema Dokumentenmanagement erläutert und an Hand des Dokumentenlebenszyklus der Einsatz eines Dokumentenmanagementsystems und dessen Komponenten verdeutlicht. Die Beschreibung von drei praktischen Beispielen soll Ihnen den Einsatz, die Anforderungen an solche Systeme und den Aufbau näher bringen.

Karlsruhe Institute of Technology

Das KIT ist eine Kooperation aus dem Forschungszentrum Karlsruhe und der Universität Karlsruhe und ist damit die größte Wissenschaftseinrichtung in Deutschland.
Auf zwei Campen arbeiten und studieren ca. 8000 Wissenschaftler, Mitarbeiter und 18000 Studenten mit einem Jahresbudget von ca. 0,5 Mrd. €.

Das Steinbuch Centre for Computing ist der Zusammenschluss des Institut für wissenschaftliches Rechnen (IWR) des Forschungszentrums Karlsruhes und dem Rechenzentrum der Universität (RZ).
Das SCC wurde nach Karl Steinbuch benannt, einem Pionier der deutschen Informatik und auf dem Gebiet lernfähiger Maschinen.

Das SCC betreibt für den Oracle UCM-Dienst 9 produktive UCM-Server und 4 Server für die Inbound Refinerys, sowie 2 Test- bzw. Schulungsserver. Eine Inbound Refinery bietet Unterstützung für die Konvertierung von Dokumenten und für die Erstellung von Miniaturbildern.
Zu unseren Kunden gehören KIT-Institute wie die Rechtsabteilung, die Dienstleistungseinheit Innovationsmanagement, die Stabsabteilung Presse, Kommunikation und Marketing und das Institut KIT-Sicherheitsmanagement. Insgesamt werden momentan im KIT mit Hilfe des Dokumentenmanagementsystems über 35000 Dokumente verwaltet.
Für die Neuentwicklung, Weiterentwicklung und Optimierung dieser Systeme arbeiten wir bereits seit mehreren Jahren erfolgreich mit der Karlsruher Firma virtual7 GmbH zusammen. Diese ist auf Oracle basierte Portal- und Middleware-Lösungen spezialisiert.

Begrifflichkeiten im Oracle UCM

Unter Dokumentenmanagement werden Methoden, Anweisungen und Prozesse für die Lenkung, Ablage und Verwaltung von Dokumenten verstanden.
Dies kann sowohl in elektronischer Form, zum Beispiel in einem Dokumentenmanagement- bzw. Contentmanagement-System, als auch klassisch in Papierform geschehen.

In einem Dokumentenmanagementsystem werden die Dokumente (Textdokumente, Zeichnungen, Bilder, AV-Medien etc.) und die dazu gehörenden, beschreibenden Informationen (Metadaten) zentral gespeichert und den Mitarbeitern, je nach Befugnis, zugänglich gemacht. Ein Dokument und seine Metadaten bilden dabei eine Einheit. Entweder werden Dokument und Metadaten zusammen in einer Datenbank gespeichert oder es werden nur die Metadaten in der Datenbank gespeichert und das dazugehörige Dokument wird im Filesystem abgelegt und referenziert.

Wesentliche Eigenschaften eines Dokumentenmanagementsystems sind visualisierte Ordnungsstrukturen, Checkin/Checkout, Versionierung sowie datenbankgestützte Metadatenverwaltung zur Index-gestützten Dokumentensuche. So gekennzeichnete Dokumente sind über mehr Informationsfelder recherchierbar, als sie ein Dateisystem zur Verfügung stellt. Im Dateisystem kann der Anwender nur über Dateiname, ggf. Dateiendung, Größe oder Änderungsdatum suchen. Beim Dokumentenmanagement stehen beliebige Felder zur Verfügung wie z. B. Kundennummer, Auftragsnummer, Betreuer usw. Eine wesentliche Anwendung des Dokumentenmanagements im engeren Sinn ist die elektronische Akte, in der aus verschiedenen Quellen alle zusammengehörigen Informationen zusammengeführt werden.

Lebenszyklus des Content

Der Dokumentenlebenszyklus beschreibt den Prozess, den ein Dokument von der Erstellung über die Verwendung bis zum Löschen durchläuft. All diese Phasen können durch ein Dokumentenmanagementsystem unterstütz werden.

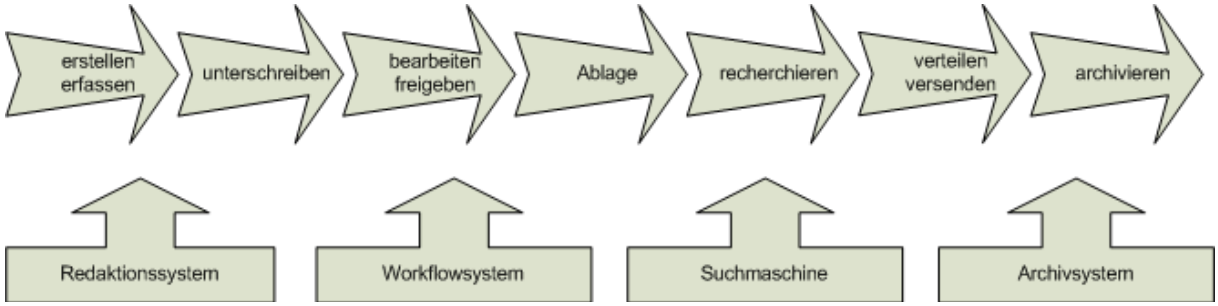

Abb. 1: Lebenszyklus des Content

Die Abbildung 1 zeigt einen möglichen Dokumentenlebenszyklus. Je nach Anforderung kann es hiervon auch Abweichungen geben.
Die Hauptaufgabe des Dokumentenmanagementsystems liegt in der Abbildung, Steuerung und Überwachung dieser Prozess über das System. Es besteht in der Regel aus mehreren Komponenten.

Mit Hilfe eines Redaktionssystems gelangen Dokumente bzw. Inhalte in das System. Dies kann durch Autorensysteme, wie Textverarbeitungsprogramme, oder auch über Scanner geschehen.

Mit dem Workflowsystem werden die eigentlichen Arbeits- und Systemprozesse abgebildet. Es übernimmt den Transfer der Informationen zwischen den Systemteilnehmern, Überwacht die Ausführung und warnt beispielsweise, wenn Bearbeitungszeiten überschritten werden.
Mit einer Suchmaschine kann auf Inhalte aus der Ablage zugegriffen werden. Standardmäßig ist eine Volltextsuche sowie eine Suche nach Metadaten möglich.
Mit einem Archivsystem werden Dokumente, die nicht mehr in Bearbeitung oder nicht mehr im Prozess sind, archiviert. Diese können nicht mehr verändert werden und werden erst gelöscht, wenn ihre Aufbewahrungsfrist abgelaufen ist. Das Archiv stellt somit die letzte Station im Dokumenten-lebenszyklus dar.

Einsatz im KIT

Die Vorteile für den Einsatz von Oracle UCM für uns und unsere Kunden ist die Beschleunigung der Prozesse, der Zugriff ist unabhängig von Ort und Zeit, die Dokumente sind per Mausklick am Monitor verfügbar und mehrere Mitarbeiter können gleichzeitig von verschiedenen Standorten auf die Dokumente zugreifen. Die Dokumente werden strukturiert, revisionssicher und redundanzfrei abgelegt und können Archiviert werden.

Dokumentenmanagementsystem

Am KIT wurde mit Oracle UCM ein Dokumentenmanagementsystem für die Ablage von Vertragsblättern für die Rechtsabteilung entwickelt.

Der Mitarbeiter scannt die Dokumente ein. Der Scanner sendet dieses als tiff an den ScannRouter-Server. Auf diesem Server wandelt das Programm CVista PDF Compressor das tiff-Dokument in eine PDF um. Dabei erfolgt automatisch die OCR, die optische Zeichenerkennung. Dies ist ein Verfahren um Abbildungen von Texten, z.B. ein eingescannter Brief, wieder in binäre Textinformationen zurückzuwandeln.
Dieses Verfahren ist notwendig um z.B. nach bestimmten Schlüsselwörtern im eingescannten Dokument suchen zu können. Auch automatische Überprüfungen von z.B. Rechnungen sind damit möglich oder die automatische Identifikation von Dokumenten, handelt es sich um eine Rechnung oder eine Bestellung.

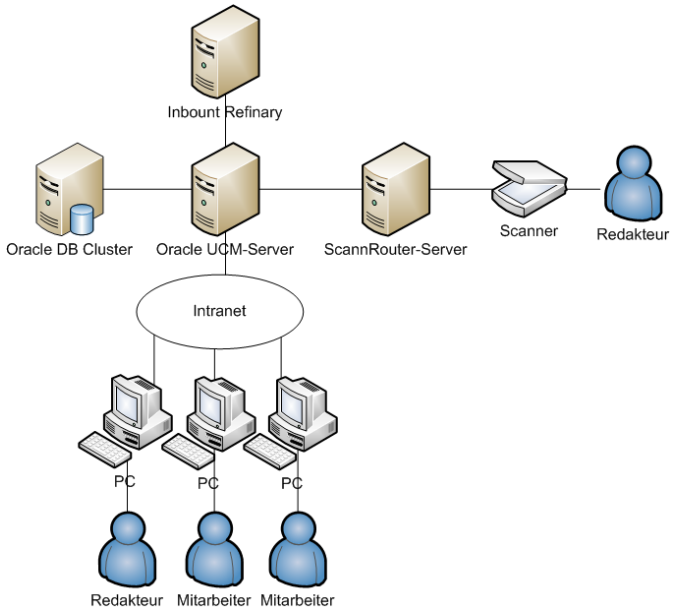

Abb. 2: Dokumentenmanagementsystem am KIT

Auf dem Oracle UCM-Server ist eine BatchLoader-Komponente definiert. Dieser holt das auf dem ScannRouter-Server umgewandelte PDF-Dokument ab und checkt es automatisch mit vordefinierten Metadaten in das Dokumentenmanagementsystem ein.

Beim Eincheckvorgang wird das PDF zunächst zur Inbound Refinery gesandt. Die Inbound Refinery erstellt bestimmte Formate vom eingecheckten Dokumente. Aus Word-Dokumenten können z.B. PDFs oder HTML-Dateien erzeugt, oder aus TIFF-Bildern können kleinere PNGs, JPEGs erstellt werden.

Parallel dazu werden die angegebenen Metadaten in die Datenbank geschrieben, in unserem Fall eine Oracle Datenbank in einem Datenbank Cluster. Die Dokumente selber werden zusammen mit den erstellten Formaten in bestimmte Ordner auf dem Oracle UCM-Server abgelegt.

War der Eincheckvorgang erfolgreich, kann der Mitarbeiter das Dokument im Administrationsportal finden, die Metadaten weiter online im Browser bearbeiten oder es einer Sammelmappe zuordnen.

Eine Sammelmappe ist eine eigenentwickelte Komponente, um Dokumente in einer Art Aktentasche zusammenzufassen. Mehrere Dokumente gleicher Merkmale können somit gruppiert und geordnet werden. In unserem Fall werden mehrere Vertragsdokumente von einem Vertragspartner einer Sammelmappe zugeordnet. Die entsprechenden gemeinsamen Metadaten, wie Name des Vertragspartners, Anschrift, Ansprechpartner, Kontaktmail usw., gehören der Sammelmappe und brauchen nicht redundant den einzelnen Vertragsdokumenten zugewiesen werden.

Contentmanagementsystem

Für die Abteilung KIT-Sicherheitsmanagement wurde ein Web-Portal basierend auf Oracle UCM entwickelt. Es entspricht einer Oracle UCM Standardapplikation mit der Komponente Site Studio.
Mit Hilfe dieser Komponente können Mitarbeiter ohne Kenntnisse in Webprogrammierung einfache Inhalte in das Portal einpflegen, ändern oder wieder löschen.
Nach dem Prinzip von WYSIWYG („**w**hat **y**ou **s**ee **i**s **w**hat **y**ou **g**et“) werden Inhalte direkt in einem kleinen Editor im Browser eingestellt. Das Ergebnis wird dem Anwender sofort angezeigt.

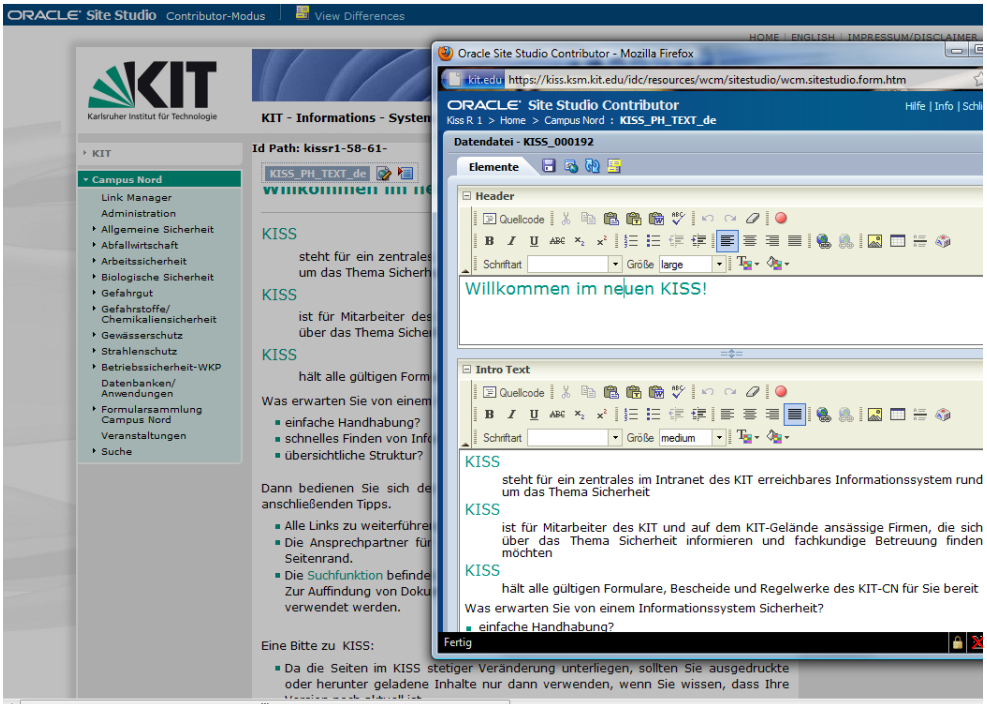

Abb. 3: KISS-Portal mit dem Site Studio Editor

Eine besondere Anforderung, die der Kunde an das Portal gestellt hatte, war es, Links in einem Linkmanager zu verwalten. Der Vorteil hierbei ist, dass diese Links, wenn sie sich ändern, nur im Linkmanager angepasst werden müssen und Redakteure nicht auf allen Webseiten nach zu ändernden Links suchen müssen.

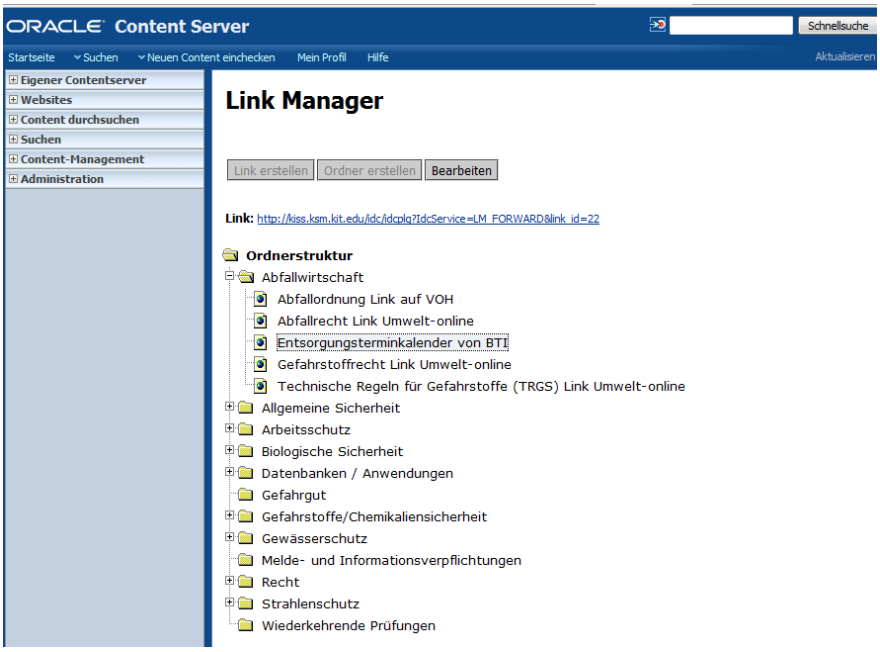

Abb. 4: LinkManager des KISS-Portals

Diese zusätzliche Komponente wurde von der Firma virtual7 GmbH entwickelt und in unserem Portal integriert. Der Redakteur kann nun in bestimmten Ordnern durch die Eingabe einer URL und eines Beschreibungstextes ein neuen Link anlegen. Das Portal erzeugt damit einen eigenen Link, hier *http://kiss.ksm.kit.edu/idc/idcplg?IdcService=LM_FORWARD&link_id=22*. Durch die *link_id* wird der Link eindeutig identifiziert.
Im Portal selbst wird lediglich der erzeugte Link veröffentlicht. Dieser verweist schließlich zur eigentlichen URL und leitet den Benutzer des Portals dorthin um.

Medienserver

Für die Ablösung einer alten Datenbank für Bilder wurde mit Hilfe von Oracle UCM eine neue Bilddatenbank aufgebaut.
Im Mittelpunkt stand hierbei die Integration der Bilder aus der alten Bilddatenbank, sowie das Einchecken neuere mit einer digitalen Spiegelreflexkamera aufgenommenen Fotos.

Die Bilddatenbank besteht aus dem Oracle UCM Standardadministrationsportal und einem Webportal. Im Administrationsportal können die Mitarbeiter Fotos und Bilder einchecken und Metadaten bearbeiten. Das Webportal ermöglicht die Recherche nach Fotos zum Thema KIT und Forschung für Mitarbeiter und externe Kunden.

Das Webportal besteht aus einer benutzerfreundlichen und intuitiven Suchseite, einer Ergebnislistenseite, die die Suchergebnisse anzeigt, und einer Detailansichtseite.

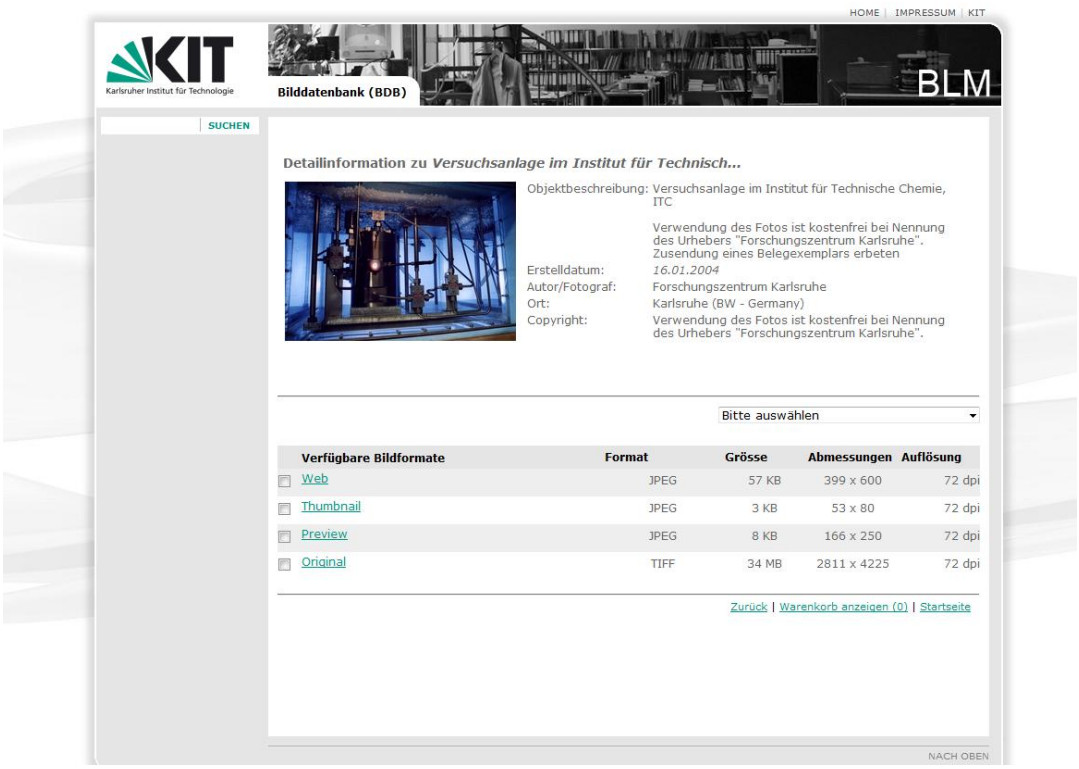

Abb. 5: Detailansichtseite der Bilddatenbank des KITs

Auf der Detailansichtseite werden zum einen das Foto zusammen mit den wichtigsten Metadaten gezeigt und zum anderen die möglichen Bildformate wie Web, Thumbnail, Preview oder das Original angegeben.

Zusätzlich wurde eine Warenkorbkomponente entwickelt, die es ermöglicht bestimmte Bildformate von verschiedenen Bildern in einem Warenkorb abzulegen und diese aus dem Warenkorb auf die lokale Festplatte zu downloaden.

Bildformate der Bilddatenbank

Die verschiedenen Bildformate werden für die Anzeige im Web benötigt. Der Grund ist, dass das eingecheckte originale Bild meistens eine sehr große Auflösung und Abmessung besitzt und in einem Format vorliegt, welches nicht geeignet ist, um es im Internet darzustellen.
Hierfür bietet Oracle UCM ein externes Programm –ImageAlchemy– an, um direkt nach dem Eincheckvorgang aus dem originalen Bild bestimmte Formate zu erstellen. ImageAlchemy ist ein konsolenbasiertes Bildkonvertierungsprogramm, mit dem die eingecheckten Bilder auf der Inbound Refinery in eine Preview-, Web- und Thumbnailversion konvertiert werden. Die dazugehörigen ImageAlchemy-Befehle sind auf dem Inbound Refinery-Server in der Konfigurationsdatei *…\custom\DAMConverter\damconverter_basedeinitions.hds* definiert.

In der Praxis traten allerdings Probleme mit ImageAlchemy auf. Insbesondere bei der Konvertierung von TIFF-Bildern im CYMK-Farbschema zu JPG-Bildern im RGB-Farbschema traten teilweise schwere Farbfehler auf. Die konvertierten Bilder waren insgesamt zu dunkel und die Farbwerte stimmten mit dem Original nicht überein.

Die Ursache liegt in der veralteten ImageAlchemy-Software von der Firma Handmade Software, die von Oracle(bis 10gR3) mitgeliefert wird.

Abb. 6: links das korrekt konvertierte und rechts das farbverfälschte Thumbnail

Um das Problem zu beheben, wurde ImageAlchemy durch das freie und kommandozeilenbasierte Graphikprogramm ImageMagick ersetzt, mit dem diese Probleme nicht mehr in dem Maße auftraten. Dafür musste das Programm ImageMagick auf dem Inbound Refinery-Server installiert und in der Konfigurationsdatei *damconverter_basedifinitions.hda* die Konvertierungsbefehle ersetzt werden.

```
Imagemagick Thumbnail
-resample 72x72 –colorspace RGB –compress JPEG –resize 80x80> -quality 50
```

Mit diesem Befehl erstellt die Inbound Refinery eine Thumbnailversion des eingecheckten Bildes. Mit resample wird die Auflösung auf 72dpi gesetzt, mit colorspace wird das Farbschema RGB festgelegt, compress definiert das Bildformat JEPG und resize legt die Bildabmessung auf 80 mal 80 Pixel fest. Der Befehl quality verringert die Bildqualität um die Größe des Bildes für die Browserdarstellung zu verringern.

Metadaten der Bilddatenbank

Die Bilder aus der alten Bilddatenbank besaßen bereits Metadaten, diese mussten lediglich in die neue Bilddatenbank integriert werden. Neuere Bilder, die ausschließlich von einem Fotographen des KITs stammen, bringen die meisten Metadaten in Form von IPTC-Daten schon mit.

Der IPTC-NAA-Standard dient der Speicherung von Informationen zu Bildinhalten in Bilddateien und definiert somit diese Metadaten. Das heißt, der Fotograph beschreibt mit diesen Metadaten seine Bilder. Das erlaubt es, Hinweise zu den Bildrechten, den Namen des Autors, Titel oder Schlagworte anzugeben und auch direkt in der Bilddatei zu speichern. Diese Art der Speicherung von Metadaten ist für Bildagenturen oder Bildarchiven zum Teil notwendig.

Die KIT-Bilddatenbank nutzt diese IPTC-Daten, ließt diese beim Einchecken aus dem Bild aus und speichert sie in einem bestimmten Metadatenfeld der Bilddatenbank ab.

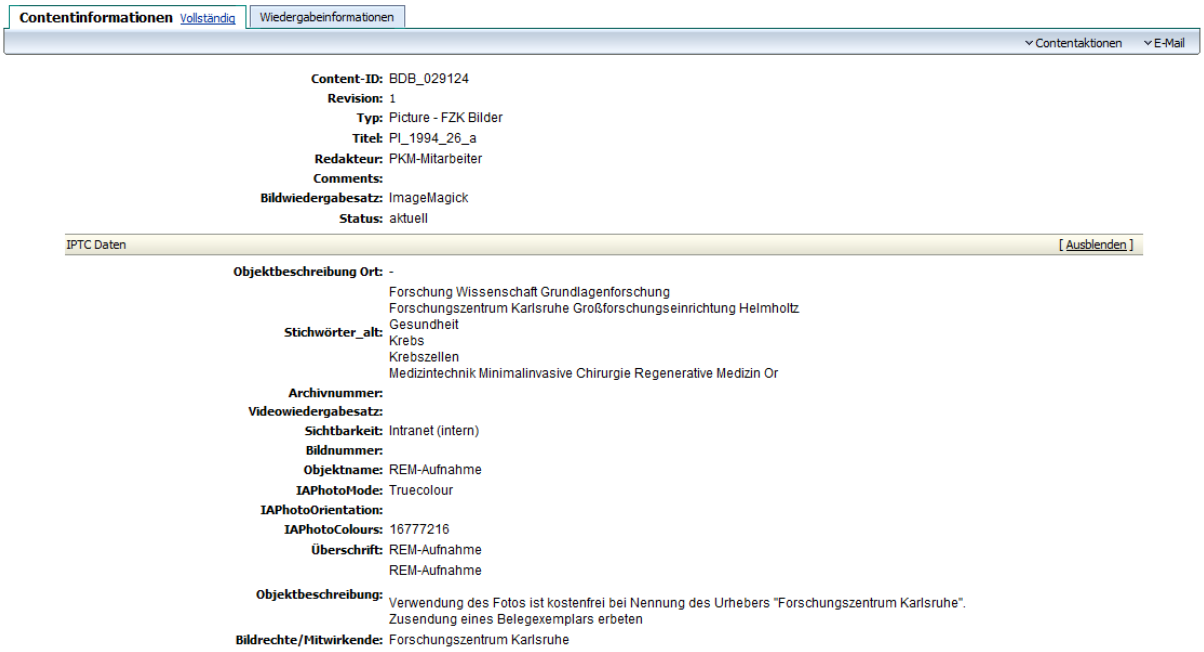

Abb. 7: Metadaten eines Bildes aus der Bilddatenbank des KITs

Neben den Standardmetadaten, wie Content-ID, Revision, Typ usw. enthält jedes Bild ein weiteres Metadatenfeld namens IPTC Daten. Dort erscheinen die aus dem Bild ausgelesenen IPTC-Daten wie Stichwörter, Objektname, Überschrift usw.
Um auch die Bilder aus der alten Bilddatenbank zu integrieren, wurde ein weiteres Metadatenfeld angelegt. Hier werden die Metadaten aus der alten Bilddatenbank gespeichert.

Berechtigungskonzept der Bilddatenbank

Für die Bilddatenbank wurde ein dreistufiges Berechtigungskonzept entwickelt. Es gibt Bilder, die dürfen im Internet von jedem gesehen werden, die von jedem Mitarbeiter im Intranet und die nur von bestimmten Personen einer Abteilung gesehen werden dürfen.

Um dies zu realisieren, wurde das Berechtigungskonzept mit Hilfe von Konten und Sicherheitsgruppen entwickelt. Dabei werden die Bilder bestimmten Konten zugeordnet, die die Zugehörigkeit des Bildes bestimmen.

z.B.:

Insitut/BDB/
Institut/BDB/INTRANET
Institut/BDB/INTERNET

Der Zugriff auf die Bilder bestimmen die Sicherheitsgruppen, *public* und *secure*. Die Kombination aus Sicherheitsgruppe und Kontozugehörigkeit regelt den Zugriff auf das Dokument. Ein Dokument mit dem Konto SCC/BDB/INTERNET und der Sicherheitsgruppe *public* kann von jedem im Internet gesehen werden. Dagegen kann ein Dokument mit dem Konto SCC/BDB/INTRANET und der Sicherheitsgruppe *secure* nur von angemeldeten Benutzern im Intranet betrachtet werden.

Kontaktadresse:

Ronny Wörl
Karlsruher Institut für Technologie
Hermann-von-Helmholtz-Platz, 1
D-76344 Eggenstein-Leopoldshafen

Telefon:	+49 (0) 7247-82 8627
Fax:	+49 (0) 7247-82 4972
E-Mail	ronny.woerl@kit.edu
Internet:	www.kit.edu