Fachhochschul-Diplomstudiengang für Telekommunikation und Medien

„NewsML by APA“ für kurier.at

Wie kann ein Nachrichtenartikel automatisch übernommen werden, um so schnell als möglich zur weiteren Verarbeitung zur Verfügung zu stehen und online zu erscheinen?

Wien, im September 2006

Diplomarbeit

„NewsML by APA' für kurier.at"

Ausgeführt zum Zweck der Erlangung des akademischen Grades eines
Dipl.-Ing. (FH) Telekommunikation und Medien
am Fachhochschul-Diplomstudiengang
Telekommunikation und Medien St. Pölten

unter der Erstbetreuung von

Bernhard Zwischenbrugger

Zweitbegutachtung von

Mag. Thomas Jöchler

ausgeführt von

Bettina Tuller tm021122

St. Pölten, am 8. September 2006 Unterschrift:

Ehrenwörtliche Erklärung

Ich versichere, dass

- ich diese Diplomarbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfe bedient habe.

- ich dieses Diplomarbeitsthema bisher weder im Inland noch im Ausland einem Begutachter/einer Begutachterin zur Beurteilung oder in irgendeiner Form als Prüfungsarbeit vorgelegt habe.

Diese Arbeit stimmt mit der von den Begutachtern beurteilten Arbeit überein.

....

Ort, Datum Unterschrift

Kurzfassung

KURIER Online bekommt Nachrichtenmeldungen aus mehreren Quellen, großteils jedoch aus einer – der APA (Austria Presse Agentur). Sie deckt mit ihren Informationen den europäischen Nachrichtenbereich fast zur Gänze ab, da sie selbst von der DPA (Deutsche Presse Agentur) und der EPA (European Press Agency) ihre Meldungen bekommt und weiter verarbeitet. Zusätzlich bezieht der KURIER noch Nachrichtenmeldungen von der AP (Associated Press) und von Reuters. Weiters werden für KURIER Online auch Inhalte, die aus der Zeitungsredaktion des KURIER kommen, verarbeitet.

Die Online-Redakteure verwenden hauptsächlich den von der APA zur Verfügung gestellten APA-OnlineManager, welcher es ihnen ermöglicht über ein benutzerfreundliches User-Interface direkt auf die Datenbank der APA zuzugreifen. Um den Workflow zu optimieren, erhält KURIER Online den Nachrichtenbasisdienst „NewsML by APA“, der direkt ins Content-Management-System (CMS) eingebunden werden soll.

Gegenstand dieser Arbeit ist die Entwicklung der Schnittstelle „NewsML zu CMS“, ohne auf die vorherrschende Umgebung Einfluss zu nehmen. Neben der Implementierung der erhaltenen Artikel im NewsML-Format, sollte auch die Darstellung und weitere Verarbeitung umgesetzt werden. So haben die Online-Redakteure die Möglichkeit einer Voransicht und auch der direkten Übernahme in das User-Interface des KURIER Online CMS, mit welchem die Artikel dann normal weiter verarbeitet werden.

Abstract

KURIER Online receives news items from several sources, mainly however from one - the APA (Austria Presse Agentur). APA covers the European news almost entirely as it obtains information from the DPA (Deutsche Presse Agentur) and the EPA (European Press Agency). In addition it receives news stories from AP (Associated Press) and Reuters. KURIER Online furthermore makes use of contents stemming from the editorial staff of the KURIER newspaper itself.

The online editors mainly use the APA-OnlineManager, a tool which allows them to gain access to the APA database. In order to optimize the workflow, KURIER Online wants to make use of a basic message service, "NewsML by APA", which is to be merged directly with the content management system (CMS) of KURIER Online.

The subject of this paper is the development of the interface "NewsML to CMS" without influencing the prevailing environment. Apart from the implementation of the received articles in NewsML format, the representation and further processing should be converted, too. The online editors should have the possibility of a preview and also be able to transfer the news item directly into the user interface of the KURIER Online CMS, where it can subsequently be processed like any other article.

1. Inhaltsverzeichnis

1. Inhaltsverzeichnis ..... - 6 -
2. Einleitung ..... - 8 -
3. NewsML – der neue Nachrichtenstandard ..... - 10 -
3.1 Warum NewsML? ..... - 11 -
3.1.1 IPTC 7901 ..... - 12 -
3.1.2 NITF ..... - 13 -
3.1.3 PRISM ..... - 14 -
3.1.4 RSS und Atom ..... - 15 -
3.1.5 IIM ..... - 16 -
4. Struktur und Aufbau ..... - 17 -
4.1 XML ..... - 17 -
4.2 XML-Schema ..... - 19 -
4.3 NewsML-Struktur ..... - 20 -
5. „NewsML by APA“ ..... - 25 -
5.1 Struktur ..... - 25 -
5.1.1 Catalog, NewsEnvelope, NewsItem ..... - 26 -
5.1.2 Identification ..... - 27 -
5.1.3 NewsManagement ..... - 29 -
5.1.4 NewsComponent ..... - 30 -
5.1.5 NewsLines ..... - 31 -
5.1.6 AdministrativeMetadata ..... - 31 -
5.1.7 DescriptiveMetadata ..... - 32 -
5.1.8 ContentItem ..... - 33 -
5.2 Vererbung von Metadaten ..... - 34 -
6. Integration von „NewsML by APA“ ..... - 37 -
6.1 Einführung in das KURIER-CMS ..... - 38 -
6.2 Einbettung von NewsML in das vorhandene CMS ..... - 40 -
6.2.1 Die Theorie ..... - 45 -
6.2.2 Die Performance ..... - 48 -
6.2.3 Die Fehleranfälligkeit ..... - 49 -

6.2.4 Die Implementierung ..... - 50 -
6.2.4.1 Die Serverumgebung ..... - 51 -
6.2.4.2 Das Layout ..... - 51 -
6.2.4.3 functions.inc.php ..... - 51 -
6.2.4.4 apabasis_simplexml.php ..... - 53 -
6.2.4.5 dom.php ..... - 58 -
6.2.4.6 parsertopic.php ..... - 60 -
6.2.4.7 apabasis.php ..... - 61 -
6.2.4.8 router.php ..... - 67 -
6.2.5 Diagramme ..... - 70 -
7. Ausblick ..... - 72 -
8. Zusammenfassung ..... - 74 -
9. Quellenverzeichnis ..... - 76 -
10. Glossar ..... - 77 -
11. Abbildungsverzeichnis ..... - 82 -
Anhang A: Quelltexte ..... - 83 -
A.1 functions.inc.php ..... - 83 -
A.2 apabasis_simplexml.php ..... - 84 -
A.3 dom.php ..... - 87 -
A.4 parsertopic.php ..... - 89 -
A.5 apabasis.php ..... - 89 -
A.6 router.php ..... - 94 -
Anhang B: NewsML ..... - 97 -
B.1 NewsML von Reuters ..... - 97 -
B.2 NewsML von AFP (Agence France Presse) ..... - 101 -
B.3 „NewsML by APA“ ..... - 103 -

2. Einleitung

Die optimale Versorgung von Agenturen und Redaktionen mit Nachrichten ist gerade im Informationszeitalter ein wichtiger Punkt. Wer bringt die aktuellsten Nachrichten schneller online? Hier stellt die Geschwindigkeit der Übertragung und Verarbeitung der Daten den entscheidenden Faktor dar. Bis jetzt haben die Redakteure von KURIER Online über den APA-OnlineManager (AOM) Zugang zu aktuellen Daten und übertragen sie per „Copy / Paste“ ins vorhandene CMS. Der Artikel wird überarbeitet und exportiert, dabei werden die eingegeben Daten in die Datenbank übernommen. Sobald der Artikel exportiert ist, wird er auf dem Webserver statisch abgespeichert und erscheint online.

Maßgebend bei der Verarbeitung von wichtigen Nachrichten ist wie schon erwähnt die Geschwindigkeit. Durch die Möglichkeit der direkten Übernahme der Inhalte der von der APA bereitgestellten NewsML-Dokumente ins CMS, könnten aktuelle Meldungen in eine Art „Nachrichtenpool“ übernommen werden. Aus diesem Pool könnten sich dann die Online-Redakteure Material für ihre Artikel zusammensuchen, ähnlich dem zurzeit benutzen AOM, mit dem Vorteil, dass sie sich direkt im KURIER-System befänden und ohne großen Aufwand die Stories übernehmen und bearbeiten könnten. In weiterer Folge könnten durch Priorisierung die wichtigsten Artikel automatisch übernommen werden und direkt, ohne die zeitaufwändige Über- und Verarbeitung, online gestellt werden.

Es ist ein „Error Handling“ zu berücksichtigen, das überprüft, ob die Daten vom Feed auch ins CMS übernommen wurden oder das aus irgendeinem Grund nicht möglich ist (das CMS könnte z.B. die Daten nicht richtig interpretieren auf Grund eines Fehlers im NewsML-Dokument). In diesem Fall muss dafür gesorgt sein, dass eine Benachrichtigung über den Fehler per Mail versendet wird.

Im Gegensatz zu einer statischen Website, die nur einen Standard zur Zeichenkodierung verwendet, treffen hier gleich mehrere verschiedene Kodierungen aufeinander. Es gilt also auf die korrekte Verarbeitung folgender

Elemente zu achten: zuerst die Übernahme vom NewsML-Feed ins CMS, dann vom CMS in die Datenbank und schließlich von der Datenbank in das statische Dokument am Webserver.

Die letzten beiden Schnittstellen sollen in dieser Arbeit nicht so große Beachtung finden als die erste, da diese bereits einen funktionierenden Prozess darstellen. Insgesamt soll sichergestellt werden, dass Sonderzeichen, im Besonderen Umlaute, richtig dargestellt werden. Weiters soll auch Wert auf die hohe Performanz und die geringe Fehleranfälligkeit dieser Schnittstelle gelegt werden. Die Implementierung von „NewsML by APA“ in das CMS von KURIER Online ist Hauptgegenstand dieser Arbeit.

3. NewsML – der neue Nachrichtenstandard

Um den Austausch von Nachrichten zwischen Anbieter und Abnehmer zu vereinfachen, wurde im Oktober 2000 von der IPTC (International Press Telecommunications Council) ein neuer Standard verabschiedet – NewsML. Im Jahre 1998 hatte die Nachrichtenagentur Reuters die Idee zu diesem einheitlichen Nachrichtenformat[1] und reichte das Konzept dann 1999 an die IPTC weiter.[2] Diese setzte das Format NewsML als Nachrichtenstandard um und entwickelte daraus auch weitere Formate, wie etwa SportsML oder ProgramGuideML. Diese „Verwandten" sind, wie NewsML, spezielle Formate zur Übertragung von Sportdaten, wie etwa Spielstände, Spielpläne oder Statistiken, bzw. Programmhefte für Radio oder Fernsehen.[3] NewsML hingegen ist eigens für die Übertragung von Nachrichten, im Sinne von Informationen zum Zeitgeschehen, entwickelt worden, um dem Chaos in der Übermittlung verschiedener Datentypen Einhalt zu gebieten. NewsML unterstützt also die Kommunikation zwischen Nachrichtendienststellen und Endbenutzern, zwischen Nachrichtenagenturen und deren Kunden, aber auch innerhalb und zwischen Redaktionen.[4]

NewsML basiert auf XML (Extensible Markup Language), einem *„Standard zur Erstellung maschinen- und menschenlesbarer Dokumente in Form einer Baumstruktur"*.[5] XML ist eine erweiterbare Auszeichnungssprache, was Entwicklern die Möglichkeit bietet, *„ihre eigenen Formate zum Speichern und Weitergeben von Informationen zu schaffen"*.[6]

---

[1] vgl. „Welcome to NewsML" von Reuters, online unter <http://about.reuters.com/newsml/newsml.asp> (10.06.2006)

[2,4] vgl. „NewsML – die Lösung für das Nachrichtenchaos" auf www.ipsi.fraunhofer.de, online unter <http://www.ipsi.fraunhofer.de/merit/lehre/veranstaltungen/intercosmical2006/material/InterCOSMICAL2006_3.I.1_NewsML_OezguerDagdelen.pdf> (10.06.2006)

[3] vgl. „IPTC Standards" von der IPTC, online unter <http://www.iptc.org/pages/index.php> (10.06.2006)

[5] „Extensible Markup Language" von Wikipedia, online unter <http://de.wikipedia.org/wiki/XML> (10.06.2006)

[6] Eric van der Vlist, 2003, S. 1

NewsML, als neues Format einer Markup Language, zeichnet sich dadurch aus, dass es medien- und plattformunabhängig ist.[7] Das bedeutet, dass diese angepasste Form einer XML-Struktur es nicht nur erlaubt Text zu übertragen, sondern auch dem Nachrichteninhalt zugehöriges Material wie z.B. Video, Audio oder Grafiken. Dies kann durch das Einbinden der Binärdaten geschehen oder durch simple Referenzierung und ist für mehrere Plattformen anwendbar, wie z.B. Internet, Fernsehen oder Rundfunk, aber auch im Bereich der Mobiltelefonie.

Gerade im Informationszeitalter gibt es kaum noch Unternehmen, die sich auf eine Art der Kommunikation spezialisieren, vielmehr wird die ganze multimediale Vielfalt ausgenutzt, um Nachrichten über so viele Plattformen wie nur irgendwie möglich zu verteilen.[8]

3.1 Warum NewsML?

Durch die Bündelung aller relevanten Daten in einem Dokument kann Information schneller erfasst und verarbeitet werden. Dadurch dass NewsML als Branchenstandard definiert wurde und derzeit von über 20 offiziellen Anbietern verwendet und an deren Nachrichtenabnehmer weiterverteilt wird, ist hier auch der Faktor der Kosten zu erwähnen. Die Verwaltung von Nachrichten wird dadurch nicht nur einfacher und schneller, sondern auch kostengünstiger, da NewsML bereits weltweit Anwendung findet, vor allem aber in den größten Nachrichtenagenturen in Nordamerika, Europa und Asien.

NewsML ist ein flexibles Format das einheitlich mit den unterschiedlichen Medientypen verfährt und von der verwendenden Agentur in der Grundstruktur

[7] vgl. „Introduction to NewsML“ von der IPTC, online unter <http://www.newsml.org/pages/intro_main.php> (05.06.2006)

[8] vgl. „NewsML – die Lösung für das Nachrichtenchaos“ auf www.ipsi.fraunhofer.de, online unter <http://www.ipsi.fraunhofer.de/merit/lehre/veranstaltungen/intercosmical2006/material/InterCOSMICAL2006_3.I.1_NewsML_OezguerDagdelen.pdf> (10.06.2006)

angepasst werden kann.[9] Nachrichten sind somit personalisierbar, so kann man z.B. einem einzigen Bild Texte in mehreren Sprachen zuordnen (siehe Abbildung 2 in Kapitel 4.3 NewsML-Struktur).[10] Nachrichten im NewsML-Format sind außerdem leichter zu identifizieren, da sie einen weltweit eindeutigen und einmaligen Identifikationsnamen tragen (siehe Kapitel 5.1 Struktur). Weiters führt NewsML eine Versionsnummer mit, die auf eine korrigierte oder aktualisierte Version der Nachricht hinweist. Somit ist es möglich, dass ältere Versionen durch die aktuellste ersetzt werden können. Ein weiteres Datenfeld stuft den Status der Meldung z.B. als „Usable“ oder „Canceled“ ein. NewsML bietet außerdem etwas, was bisher kein anderes Nachrichtenformat bieten konnte, nämlich die Fähigkeit mehrere Inhalte in einem Paket zu transportieren. Dabei ist es völlig egal, ob die Inhalte vom gleichen Medientyp sind oder nicht. Zusätzlich ist es durch die Möglichkeit zur Vergabe von „Rollen“ (siehe Kapitel 5.1.4 NewsComponent) möglich, die einzelnen NewsItems in Relation zueinander zu setzen. Durch NewsML ist auch eine klare Trennung des eigentlichen Inhalts von den beschreibenden Metadaten und den organisatorischen Mitteilungen gewährleistet.[11]

NewsML löste nicht nur einige Standards im Nachrichtenbereich ab, sondern kooperiert auch mit manchen davon. Andererseits gibt es natürlich auch konkurrierende Formate. In diesem Zusammenhang werden immer wieder dieselben Normen genannt.

3.1.1 IPTC 7901

Dieses Format diente zum Austausch von Textnachrichten zwischen Nachrichtenagenturen, Zeitungen und anderen Empfängern. Es wurde 1979 von der IPTC entwickelt und regelmäßig angepasst. Die letzte Version (Version 5) wurde 1995 freigegeben. Seit diesem Zeitpunkt steht diese Entwicklung still,

[9,11] vgl. „Mindestens 5 gute Gründe, NewsML einzusetzen“ von APA-IT Informations Technologie GmbH, online unter <http://www.apa-it.at/NewsML/Deutsch/GuteGruendeFuerNewsML.pdf> (11.06.2006)

[10] vgl. „NewsML – die Lösung für das Nachrichtenchaos“ auf www.ipsi.fraunhofer.de, online unter <http://www.ipsi.fraunhofer.de/merit/lehre/veranstaltungen/intercosmical2006/material/InterCOSMICAL2006_3.I.1_NewsML_OezguerDagdelen.pdf> (10.06.2006)

obwohl es noch immer in vielen Teilen der Welt verwendet wird.[12] Neben IPTC 7901 gibt es noch ein ähnliches Format, ANPA 1312, welches aber hauptsächlich in den USA verwendet wird. IPTC 7901 wird vorwiegend in Europa benutzt und wurde auch seitens der APA (zusammen mit IIM für Fotos) verwendet, bis es im Jahr 2000 durch NewsML ersetzt wurde.

Der Grund dafür war einerseits der, dass IPTC 7901 nur eine sehr eingeschränkte Möglichkeit der Identifizierung bietet, nämlich eine laufende Nummer am Beginn der Meldung. Ohne die Angabe des Aussende-Datums konnte so keine Eindeutigkeit garantiert werden. Weiters war es in IPTC 7901 nur im Text ersichtlich, ob es sich um eine aktualisierte Version handelte, wohingegen NewsML mit Versionsnummern arbeitet, die dies klar ersichtlich machen. Schließlich sprach auch der in IPTC 7901 beschränkte Umfang an Metadaten-Elementen für eine Verwendung von NewsML.[13]

3.1.2 NITF

NITF steht für „News Industry Text Format“, wurde ebenfalls von der IPTC entwickelt und gilt als Vorfahre von NewsML. Es wurde in den frühen 90er Jahren entwickelt und dann im Zuge der Entwicklung von XML im Jahre 1998 der XML-Struktur angepasst. Laut IPTC wurde 2005 die letztgültige Version 3.3 fertig gestellt.[14]

Auf den ersten Blick unterscheidet sich NITF nicht wirklich von HTML (Hypertext Markup Language). Es hat die gleichen Grundzüge wie <head> Tag und <body> Tag, allerdings hat es in diesem Anwendungsbereich gewisse Vorzüge gegenüber dem Hypertext. Mit NITF war es bereits möglich mehrere Metadaten, sprich Tags mit Informationen über den Nachrichteninhalt, einzubinden, also wer

[12] vgl. „IPTC 7901: the text transmission format“ von der IPTC, online unter <http://www.iptc.org/IPTC7901/> (14.06.2006)

[13] vgl. „Mindestens 5 gute Gründe, NewsML einzusetzen“ von APA-IT Informations Technologie GmbH, online unter <http://www.apa-it.at/NewsML/Deutsch/GuteGruendeFuerNewsML.pdf> (11.06.2006)

[14] vgl. „Introduction“ von NITF von der IPTC, online unter <http://www.nitf.org/intro.php> (14.06.2006)

z.B. das Copyright hat, wann es berichtet wurde, wer es geschrieben hat und noch einiges mehr. Der Grund warum NITF jedoch schließlich etwas ins Abseits gedrängt wurde ist jener:

„There is no concept in NewsML of a paragraph or subheadline. There is no concept in NITF of a sidebar or alternative translations of the same document." [15]

Was bedeuten soll, dass es in beiden Strukturen Bereiche gibt, die technisch nicht berücksichtigt wurden, in denen sie sich aber, wenn man NITF und NewsML zusammen nutzt, ergänzen können. So ist NITF nicht für die Übertragung von Multimediadaten geeignet, sondern wird hauptsächlich für die Beschreibung von Text verwendet. Offensichtlich ist also eine exklusive Nutzung beider Standards nicht optimal. Aus diesem Grund werden NITF und NewsML häufig zusammen verwendet.

3.1.3 PRISM

Im Jahre 1999 wurde die „PRISM Working Group" (PRISM steht hierbei für Publishing Requirements for Industry Standard Metadata) eingerichtet, um eine Form der Informationsübertragung zu deklarieren, die zumindest von dieser „Working Group" und deren Kooperationspartnern einheitlich verwendet wird. Die Mitglieder dieser Arbeitsgruppe repräsentierten sowohl Softwareanbieter als auch einige Verlage.

NewsML konnte sich gegenüber PRISM letztlich nicht durchsetzen, da es nicht den Ansprüchen der Unternehmen im Printbereich genügte. So wurde PRISM sowohl für die Verwendung im Webbereich oder auch im mobilen Bereich, als auch im Verlagswesen zugeschneidert, unter zu Hilfenahme bereits existierender Standards, wie RDF (Resource Description Framework) auf Basis von XML-

[15] vgl. „Introduction" von NITF von der IPTC, online unter <http://www.nitf.org/intro.php> (14.06.2006)

Syntax. Im April 2001 wurde die erste Version von PRISM vorgestellt, die bisher letzte Version (Version 1.3) wurde im Juli 2005 veröffentlicht.[16]

3.1.4 RSS und Atom

RSS ist ein Format das Informationen über Inhalte im Webbereich maschinen- und auch menschenlesbar darstellt. Die Abkürzung RSS hat verschiedene Bedeutungen: so wird es als Rich Site Summary, RDF Site Summary oder auch als Really Simple Syndication ausgelegt. Die älteste Version (0.90) wurde 1999 veröffentlicht und basierte auf RDF. Die nächste Version (0.91) wurde im Jahr 2000 veröffentlicht und ließ RDF erstmals außer Acht, während Version 1.0 (ebenfalls im Jahr 2000 veröffentlicht) wieder erneut auf RDF basierte. Die heute noch gebräuchlichen Versionen beschränken sich auf die Versionen 1.0 und 2.0, letztere wurde 2002 umgesetzt.

Durch die ständigen Neu- und Weiterentwicklungen von verschiedenen Entwicklern, wie UserLand, Netscape und W3C, ist ein richtiger „RSS-Dschungel“ entstanden und durch die Verwendung von Namensräumen können die RSS-Feeds noch zusätzlich erweitert werden. Um hier den Überblick nicht ganz zu verlieren, sind RSS-Feed-Validatoren besonders zu empfehlen.

Ein anderes bekanntes Format in diese Richtung ist Atom, das eine Abwandlung von RSS darstellt. Da es ebenso wie RSS den Vorteil der Plattformunabhängigkeit bietet, wird es vor allem im Bereich der Weblogs genutzt und wurde auch überwiegend von Bloggern entwickelt.

„Atom entstand aus dem Bedarf heraus, die Vorteile der verschiedenen RSS-Versionen in einem neuen, unabhängigem Format zu vereinen und um sinnvolle Bestandteile zu ergänzen. Neben RSS wird das neue Format es schwer haben,

[16] vgl. „History of PRISM“ von der IDEAlliance, online unter <http://www.prismstandard.org/resources/historyofprism.pdf> (16.06.2006)

doch die Vorteile und starke Förderer wie Blogger.com werden dem neuen Format eine Zukunft geben.“ [17]

3.1.5 IIM

Im Jahre 1990 begannen die IPTC und die NAA (Newspaper Association of America) an einem weltweit anwendbaren Standard für alle Arten von Daten zu arbeiten, woraufhin 1991 das IIM (Information Interchange Model) vorgestellt wurde. Nach dem Aufkommen neuer Technologien in diesem Bereich, vor allem mit der Veröffentlichung von XML, wurde die Entwicklung von IIM im Jahre 1997 eingestellt. Die letztgültige Version von IIM ist die Version 4.1.

IIM ist darauf ausgerichtet Objekte wie Fotos, Text oder eine Kombination aus mehreren Typen zu übertragen. Die Metadaten-Elemente von IIM enthalten Informationen zu Objekttyp und Dateiformat. Aber auch zusätzliche Informationen, wie z.B. Titel oder Kategorie werden zusammen mit dem Objekt übertragen. So können Daten mit den zugehörigen Informationen gesendet werden.[18]

Insofern weist IIM gegenüber anderen Standardisierungen die stärksten Ähnlichkeiten zu NewsML auf. Die in Bildern enthaltenen IIM-Daten können aber dadurch, dass sie nur ein Titel-Feld bieten, überhaupt keine Eindeutigkeit garantieren. Wie auch schon beim „Partnerstandard“ IPTC 7901 ist auch hier der Umfang der Metadaten zu beschränkt, was auch in diesem Fall für die Verwendung von NewsML spricht.[19]

---

[17] „Alles über RSS“ vom RSS-Verzeichnis, online unter <http://www.rss-verzeichnis.de/alles-ueber-rss.php> (22.08.2006)

[18] vgl. „Information Interchange Model IIM: the first multi-media news exchange format“ von der IPTC, online unter <http://www.iptc.org/IIM/> (17.06.2006)

[19] vgl. „Mindestens 5 gute Gründe, NewsML einzusetzen“ von APA-IT Informations Technologie GmbH, online unter <http://www.apa-it.at/NewsML/Deutsch/GuteGruendeFuerNewsML.pdf> (11.06.2006)

4. Struktur und Aufbau

NewsML ist eine Entwicklung auf Basis der XML-Syntax. Um also den Aufbau von NewsML zu verstehen, ist es sinnvoll sich dessen Grundlage anzusehen.

4.1 XML

Um mit XML zu arbeiten benötigt man so gut wie keine Programmierkenntnisse, da XML keine Programmiersprache, sondern eine Auszeichnungssprache ist, vergleichbar mit HTML. XML stellt unterschiedliche Daten strukturiert dar, was das Lesen und die Handhabung, nicht nur für den Computer, vereinfacht. Es ist plattformunabhängig und zudem auch einfach zu generieren und erweitern.[20] Im Sinne der Spezifikation des W3C (World Wide Web Consortium) muss ein XML-Dokument wohlgeformt sein, das bedeutet, dass es bestimmte Regeln zu erfüllen hat:

„Es existiert genau ein Element, genannt Wurzel oder Dokument-Element, von dem kein Teil im Inhalt eines anderen Elements enthalten ist. Für alle anderen Elemente gilt: […] Die Elemente, begrenzt durch Start- und End-Tag, sind korrekt ineinander verschachtelt. […] Als eine Konsequenz daraus gilt, dass für jedes Element k, das nicht die Wurzel ist, ein anderes Element v existiert, so dass sich k im Inhalt von v befindet, aber nicht im Inhalt eines anderen Elements, das sich im Inhalt von v befindet. v heißt Vater von k, und k heißt Kind von v.“ [21]

[20] vgl. „XML in 10 Punkten“ vom W3C, online unter <http://www.w3c.de/Misc/XML-in-10-points.html> (23.08.2006)

[21] „Extensible Markup Language (XML) 1.1“ vom W3C, online unter < http://edition-w3c.de/TR/2004/REC-xml11-20040204/> (23.08.2006)

Ein Beispiel eines XML-Dokuments um dies zu veranschaulichen (Auszug): [22]

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<library>
   <book id ="b3810518883" available="true">
      <title>Auf den Hund gekommen</title>
   </book>
</library>
```

Das Dokument beginnt mit einer XML-Deklaration, welche ebenfalls eine Regel zur Wohlgeformtheit darstellt. Das Element <library> ist in diesem Fall das Wurzelelement. Das erste und einzige Kindelement von <library> stellt das Element <book> dar. Dieses Element enthält bereits Attribute, welche nur im Start-Tag vorkommen dürfen. Das Element <book> ist aber nicht nur ein Kindelement von <library>, sondern auch das Elternelement von <title>, welches wiederum das Kindelement von <book> darstellt. Im letzten Element <title> ist nun ein Textknoten vorhanden, welcher sich zwischen dem Start- und End-Tag befindet.

Für die Beschreibung von XML kann eine Dokumenttypdefinition (DTD) verwendet werden, in welcher die benötigten Elemente und Attribute deklariert werden. Die DTD ist jedoch nicht zwingend notwendig und auch nicht die einzige Schemasprache für XML. Die Festlegung kann auch über Relax NG (Relax New Generation) erfolgen, das den Vorteil hat, dass das Schema selbst ein XML-Dokument sein kann, aber nicht muss.[23] Zusätzlich wurde vom W3C noch eine weitere Schemasprache für XML eingeführt die in Form eines XML-Dokuments vorliegt – XML-Schema (XSD).

[22] vgl. Eric van der Vlist, 2003, S. 6

[23] „Relax NG home page“ von James Clark, online unter <http://relaxng.org/> (23.08.2006)

4.2 XML-Schema

XSD hat, zumindest der DTD gegenüber, den Vorteil, dass es Text und Zahlen unterscheiden kann und insgesamt ausdrucksfähiger ist als die DTD. Um einen groben Überblick über das Aussehen eines XML Dokuments in Zusammenhang mit dessen Schema zu geben werden in diesem Kapitel ein einfaches XML-Beispiel und das dazugehörige XSD dargestellt. Ein XML-Dokument ohne XML-Schema sieht so aus (Auszug):[24]

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<library>
   <book id ="b3810518883" available="true">
      <title lang="de">Auf den Hund gekommen</title>
      <author id="CMS">
         <name>Charles M. Schulz</name>
         <born>1922-11-26</born>
         <dead>2000-02-12</dead>
      </author>
   </book>
</library>
```

Mit Hilfe des XML-Schemas werden die Elemente und Attribute definiert, in folgendem Fall nach dem W3C XML-Schema (Auszug):[25]

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="library">
  <xs:complexType>
   <xs:sequence>
    <xs:element name="book" maxOccurs="unbounded">
     <xs:complexType>
      <xs:sequence>
       <xs:element name="title">
        <xs:complexType>
         <xs:simpleContent>
          <xs:extension base="xs:string">
           <xs:attribute name="lang" type="xs:language"/>
          </xs:extension>
         </xs:simpleContent>
        </xs:complexType>
       </xs:element>
```

[24] vgl. Eric van der Vlist, 2003, S. 6

[25] vgl. Eric van der Vlist, 2003, S. 17

```
          <xs:element name="author" minOccurs="0"
          maxOccurs="unbounded">
            <xs:complexType>
              <xs:sequence>
                <xs:element name="name" type="xs:string"/>
                <xs:element name="born" type="xs:date"/>
                <xs:element name="dead" type="xs:date"/>
              </xs:sequence>
              <xs:attribute name="id" type="xs:ID"/>
            </xs:complexType>
          </xs:element>
        </xs:sequence>
        <xs:attribute name="id" type="xs:ID"/>
        <xs:attribute name="available" type="xs:boolean"/>
      </xs:complexType>
    </xs:element>
  </xs:sequence>
 </xs:complexType>
</xs:element>
</xs:schema>
```

Hier wird als erstes der Namensraum („namespace”) angegeben in welchem das XML-Vokabular deklariert ist. Weiters werden für alle Elemente und Attribute so genannte Typen vergeben und die genaue Struktur festgelegt, welche Elemente Kind-Elemente enthalten können, sowie die Auflistung der Attribute die einem Element zur Verfügung stehen. So ist der Tag <name> mit einem `string` befüllt, einer Zeichenkette und untersteht dem Element <author>. Die beiden Elemente <born> und <dead> wiederum enthalten ein Datumsformat, während sich das Attribut `available` eines Booleschen Typs bedient (wahr / falsch). Das Attribut `id` wird mit einer Identifikationsnummer versehen und `lang` gibt Aufschluss über die verwendete Sprache im <title>.

4.3 NewsML-Struktur

NewsML fungiert als „Verpackung“ von Mediadaten und kann die Inhalte unmittelbar „einpacken“ oder auf die zu verwendende Datei verweisen. Zur Ansicht ist hier ein komplettes NewsML-Dokument dargestellt. Hierbei handelt es sich um die angepasste Version der APA, die so auch von KURIER Online

verwendet wird. Auf die einzelnen Elemente wird in Kapitel 5 – „NewsML by APA“ genauer eingegangen.

```
<?xml version="1.0" encoding="windows-1252"?>
<NewsML>
 <Catalog Href="http://www.apa-it.at/NewsML/catalog/
                catalog.APAMasterCatalog.xml"/>
 <NewsEnvelope>
    <DateAndTime>20060604T035642+0200</DateAndTime>
    <NewsService FormalName="Agencies"/>
    <NewsProduct FormalName="APA Basisdienst"/>
  </NewsEnvelope>
   <NewsItem>
     <Identification>
       <NewsIdentifier>
         <ProviderId>apa.at</ProviderId>
         <DateId>20060604</DateId>
         <NewsItemId>20060604APA0017</NewsItemId>
            <RevisionId PreviousRevision="0"
            Update="N">1</RevisionId>
            <PublicIdentifier>urn:newsml:apa.at:20060604:20060604A
            PA0017:1</PublicIdentifier>
       </NewsIdentifier>
       <Label>
         <LabelType FormalName="APAIntraId"/>
         <LabelText>APA0017</LabelText>
       </Label>
        <Label>
         <LabelType FormalName="APAIntraDate"/>
         <LabelText>2006-06-04</LabelText>
       </Label>
     </Identification>
     <NewsManagement>
       <NewsItemType FormalName="News"/>
       <FirstCreated>20060604T035642+0200</FirstCreated>
      <ThisRevisionCreated>20060604T035642+0200
      </ThisRevisionCreated>
       <Status FormalName="Usable"/>
       <Urgency FormalName="5"/>
       <Instruction FormalName="fortsetzung-moeglich"/>
     </NewsManagement>
     <NewsComponent>
       <Role FormalName="Hauptinhalt"/>
       <NewsLines>
           <HeadLine>China: Militärflugzeug mit 40 Insassen
           abgestürzt - Agentur</HeadLine>
           <SubHeadLine>Utl.: Im Osten des Landes - Beamter: Fünf
           Leichen geborgen </SubHeadLine>
           <CopyrightLine>Copyright, Austria Presse Agentur, Wien,
           Österreich, Alle Rechte vorbehalten</CopyrightLine>
           <KeywordLine>Unfälle/Verkehr/Luftfahrt/China
           </KeywordLine>
       </NewsLines>
```

```
      <AdministrativeMetadata>
        <Provider><Party FormalName="APA"/></Provider>
        <Creator><Party FormalName="vos"/></Creator>
        <Source>
          <Party FormalName="APA"/>
          <Party FormalName="Reuters"/>
          <Party FormalName="AP"/>
        </Source>
      </AdministrativeMetadata>
      <DescriptiveMetadata>
         <Language FormalName="de-AT"
         Vocabulary="http://www.ietf.org/rfc/rfc3066.txt"/>
         <SubjectCode>
            <Subject FormalName="CA"
            Vocabulary="urn:newsml:apa.at:20040101:APAcat:1"/>
            <Subject FormalName="AA"
            Vocabulary="urn:newsml:apa.at:20040101:APAcat:1"/>
            <Subject FormalName="08000000"/>
            <Subject FormalName="11000000"/>
         </SubjectCode>
         <Location HowPresent="Origin">
           <Property FormalName="City" Value="Peking"/>
         </Location>
         <Property FormalName="APAstwort" Value="Unfälle"/>
         <Property FormalName="APAstwort" Value="Verkehr"/>
         <Property FormalName="APAstwort" Value="Luftfahrt"/>
         <Property FormalName="APAstwort" Value="China"/>
      </DescriptiveMetadata>
       <ContentItem>
         <MediaType FormalName="Text"/>
         <Format FormalName="XHTML"/>
         <DataContent>
           <html xmlns="http://www.w3.org/1999/xhtml">
             <head>
                 <title>China: Militärflugzeug mit 40 Insassen
                 abgestürzt - Agentur</title>
             </head>
             <body>
                 <p>In China ist einer Nachrichtenagentur zufolge
                 am Samstag ein Transportflugzeug des Militärs mit
                 mindestens 40 Menschen an Bord abgestürzt.
                 Experten aus Peking leiteten den Bergungseinsatz
                 in der Provinz Anhui, meldete die Agentur Xinhua
                 am Sonntag unter Berufung auf Militärkreise. Fünf
                 Leichen seien an der Absturzstelle im Bezirk
                 Guangde bereits geborgen worden, sagte ein
                 Beamter der Regionalbehörde. Guangde liegt 200
                 Kilometer südwestlich von Shanghai.</p>
             </body>
           </html>
         </DataContent>
       </ContentItem>
     </NewsComponent>
   </NewsItem>
</NewsML>
```

Die Darstellung dieses Dokuments in einer Baumstruktur erleichtert den Überblick über die verwendeten Elemente.

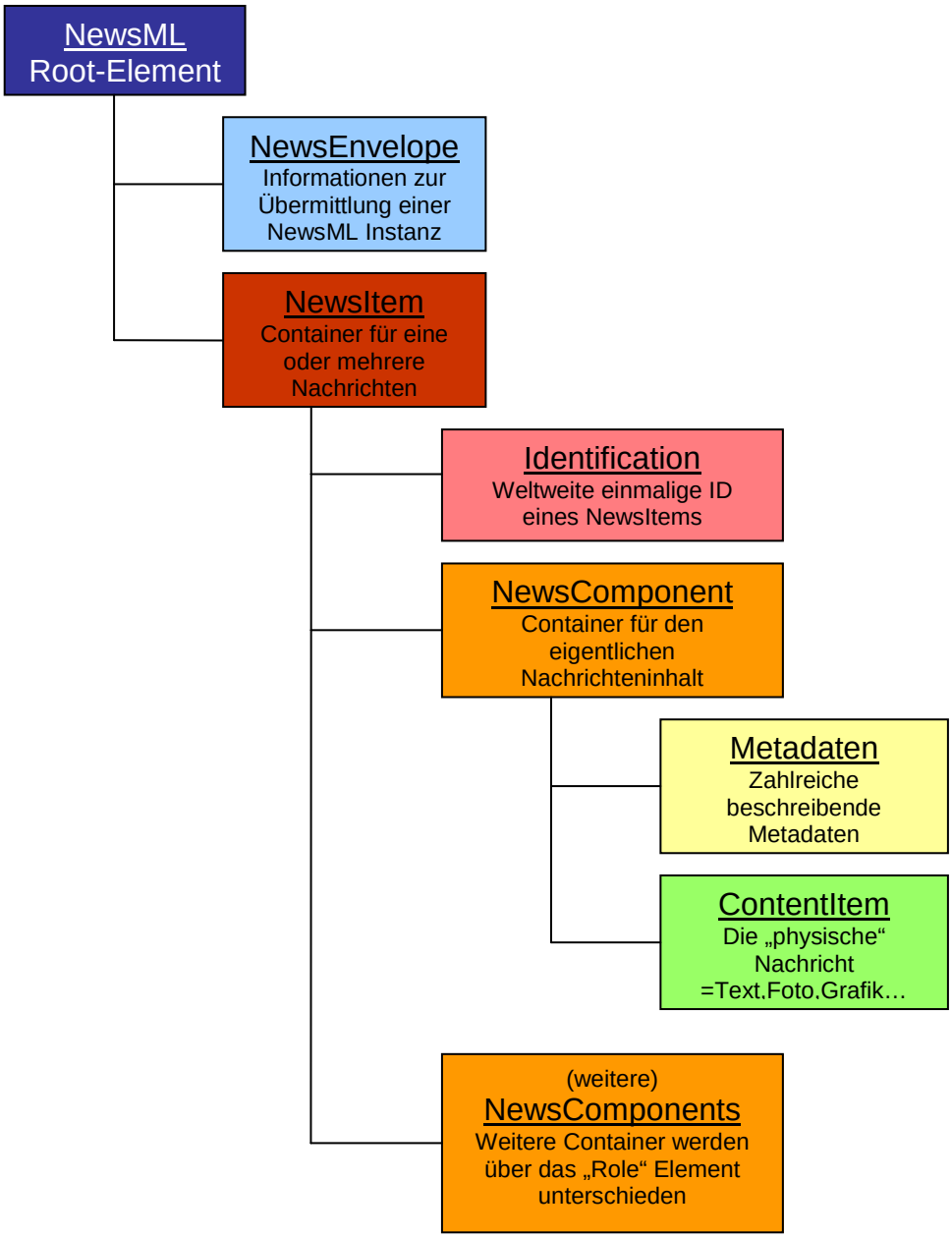

Abbildung 1: Übersicht zur NewsML-Struktur

Um genau zu veranschaulichen was in einem so genannten NewsItem alles enthalten sein kann, hier die vereinfachte Darstellung eines NewsItems mit mehreren Inhalten für mehrere Plattformen.

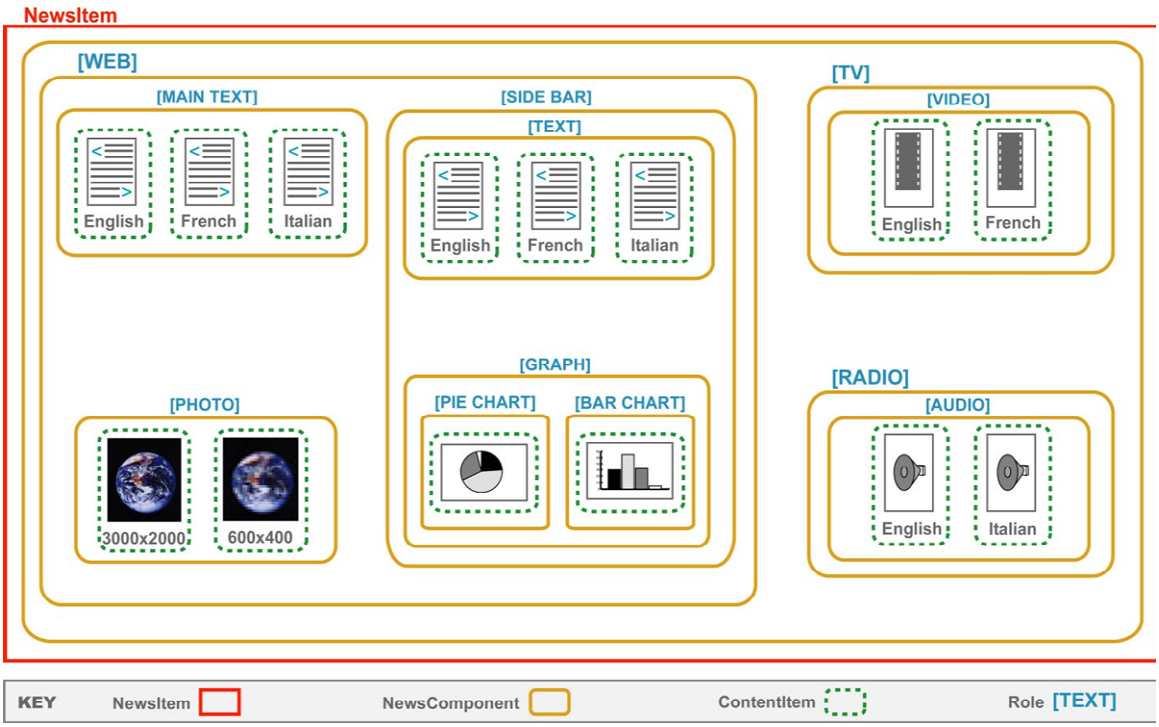

Abbildung 2: Beispielaufbau eines NewsItems

5. „NewsML by APA“

In diesem Kapitel werden die einzelnen Elemente der NewsML-Struktur sowie die Vererbung von Metadaten von „NewsML by APA“ erläutert. Wie schon erwähnt kann NewsML von den einzelnen Nachrichtenagenturen für deren Zwecke spezifisch angepasst werden. So hat dies auch die APA getan, die unter anderem KURIER Online mit NewsML beliefert.

5.1 Struktur

Der Name einer NewsML-Datei setzt sich aus dem Agenturkürzel mit laufender Nummer, sowie dem Datum der Meldung zusammen. Die von der APA gelieferten NewsML-Dokumente haben immer die Dateiendung „xml“. Unser Beispiel hat den Dateinamen „APA0017-20060604.xml“, nach dem Schema {Agenturkürzel}{ laufende Nummer für Meldungen }-{DateId}.

Das Dokument beginnt mit der XML-Deklaration, im Beispiel aus dem Kapitel 4.3 NewsML-Struktur ist dies ein XML-Header der Version 1.0. Weiters wird der Zeichensatz „Windows-1252“ verwendet.

```
<?xml version="1.0" encoding="windows-1252"?>
```

Nun beginnt die NewsML-Instanz mit dem <NewsML> Tag, dem Wurzelelement, welches drei Kindelemente enthält: <Catalog>, <NewsEnvelope> und <NewsItem>.

```
<NewsML>
   <Catalog Href="http://www.apa-it.at/NewsML/catalog/
                  catalog.APAMasterCatalog.xml"/>
   <NewsEnvelope>...</NewsEnvelope>
   <NewsItem>...</NewsItem>
</NewsML>
```

5.1.1 Catalog, NewsEnvelope, NewsItem

Im Element <Catalog> wird mit dem Attribut Href auf den „MasterCatalog“ der APA referenziert. Dieser enthält weitere Verweise zu den Vokabularen jener Elemente und Attribute, die von der APA verwaltet werden.

```
<Catalog Href="http://www.apa-it.at/NewsML/catalog/
catalog.APAMasterCatalog.xml"/>
```

Festgelegtes Vokabular gibt es zu folgenden Elementen:

- <NewsService>
- <NewsProduct>
- <Label>
- <NewsItemType>
- <Status>
- <Urgency>
- <Instruction>
- <AssociatedWith>
- <Subject>
- <Role>
- <KeywordLine>
- <Provider>
- <Source>
- <Location>
- <Property> der <Descriptive Metadata>
- <MediaType>
- <Format>

Im Element <NewsEnvelope> sind Informationen zur Übermittlung der Daten enthalten. Diese Informationen haben mit dem eigentlichen Inhalt nichts zu tun und werden (außer dem Tag <DateAndTime>) seitens des KURIER Online auch nicht ausgewertet. Hier werden die Elemente <DateAndTime>, <NewsService> und <NewsProduct> gelistet. Wie der Name bereits vermuten lässt wird im Element <DateAndTime> der Zeitpunkt der Übermittlung festgehalten. <NewsService> mit dem Attribut FormalName="Agencies" sagt aus, dass es sich um einen Agentur-Dienst der APA handelt. Hierbei muss erwähnt werden, dass die APA verschiedene Nachrichtendienste zur Verfügung

stellt, so z.B. auch „Lines“, welches ein web-orientierter Dienst ist, oder aber auch „MediaServices“, welcher von „End-Kunden-Medien“ bezogen wird, wie Radio, Fernsehen und Zeitungen.

Im <NewsProduct> wird die Art des Dienstes angegeben, in diesem Fall ist das der APA-Basisdienst. Hier könnten aber auch andere Dienste, wie z.B. der Bilderdienst oder das Original-Text-Service (OTS), ausgewiesen sein.[26]

```
<NewsEnvelope>
   <DateAndTime>20060604T035642+0200</DateAndTime>
   <NewsService FormalName="Agencies"/>
   <NewsProduct FormalName="APA Basisdienst"/>
</NewsEnvelope>
```

Das letzte Subelement nennt sich <NewsItem> und enthält die eigentliche Nachricht, sowie die dazugehörigen Metadaten. Diese können in die Bereiche <Identification>, <NewsManagement> und <NewsComponent> eingeteilt werden.

```
<NewsItem>
   <Identification>...</Identification>
   <NewsManagement>...</NewsManagement>
   <NewsComponent>...</NewsComponent>
</NewsItem>
```

5.1.2 Identification

Wie der Name schon ausdrückt werden hier alle Elemente zur Identifizierung der Nachricht gelistet. Hierfür gibt es zwei Elemente: <NewsIdentifier> und <Label>. Im <NewsIdentifier> werden Vertreiber (<ProviderId>), Datum (<DateId>), die weltweit einmalige <NewsItemId>, eine <RevisionId> (Versionsnummer) und der <PublicIdentifier> angegeben. Dieser setzt sich aus folgenden Ausdrücken zusammen:
urn:newsml:{ProviderId}:{DateId}:{NewsItemId}:{RevisionId}.

[26] vgl. „NewsML by APA: Technische Dokumentation“ von der APA-IT, online unter <http://www.apa-it.at/NewsML/Deutsch/NewsMLbyAPA-TechDokumentation_v1.06.pdf> (11.06.2006)

Die Struktur der NewsItemId weist Ähnlichkeit zum Dateinamen auf: {DateId}{Agenturkürzel}{laufende Nummer für Meldungen}. Die <RevisionId> enthält eine Zahl, die Aufschluss über erfolgte Aktualisierungen gibt. Diese Zahl wird bei Korrekturen aufwärts gezählt, so hätte eine einmal aktualisierte Meldung

den Wert 2, das Attribut PreviousRevision würde dann auf 1 gesetzt werden. Das Attribut Update gibt rudimentär darüber Auskunft, was geändert wurde. Der Wert N steht dabei für eine Aktualisierung von Inhalt und Metadaten, während der Wert A nur für eine Aktualisierung von Metadaten steht.

```
<NewsIdentifier>
   <ProviderId>apa.at</ProviderId>
   <DateId>20060604</DateId>
   <NewsItemId>20060604APA0017</NewsItemId>
   <RevisionId PreviousRevision="0" Update="N">1</RevisionId>
   <PublicIdentifier>urn:newsml:apa.at:20060604:20060604APA0017:1
   </PublicIdentifier>
</NewsIdentifier>
```

Das Element <Label> kann mehrmals vorkommen. So auch in unserem Beispiel, in dem es mit dem Subelement <LabelType> und dem Attribut FormalName einmal die APA-interne ID der Nachricht enthält, sowie ein anderes Mal das Datum der Meldung. Steht das aktuelle NewsItem in Verbindung mit einem anderen NewsItem (siehe nächsten Abschnitt AssociatedWith), werden hier für beide NewsItems Labels angegeben. Dies ist aber auch der Fall, wenn das Dokument aktualisiert wurde.

```
<Label>
   <LabelType FormalName="APAIntraId"/>
   <LabelText>APA0017</LabelText>
</Label>
<Label>
   <LabelType FormalName="APAIntraDate"/>
   <LabelText>2006-06-04</LabelText>
</Label>
```

5.1.3 NewsManagement

Im Element <NewsManagement> sind Informationen über die Nachricht enthalten, welche die Nachricht „managebar“ machen. Hier werden unter anderem der Typ des NewsItems (<NewsItemType>), der Zeitpunkt der Erstellung (<FirstCreated>), der Zeitpunkt an dem die aktuelle Version erstellt wurde (<ThisRevisionCreated>), der Status sowie Priorität (<Urgency>) der Meldung und die <Instruction> zur Weiterverarbeitung angegeben. Weiters ist es möglich mittels des Elements <AssociatedWith>, das aktuelle NewsItem mit anderen NewsItems in Verbindung zu bringen.

Der <NewsItemType> enthält in unserem Beispiel den Typ „News“. Hier könnten aber auch andere Typen, wie z.B. „Alert“, „Data“ oder „Document“ stehen. Der <Status> der Meldung kann neben dem Wert „Usable“ ebenso die Werte „Canceled“, „Withheld“ oder „Embargoed“ enthalten. Somit kann also genau angegeben werden ob die Nachricht brauchbar ist, zurückgezogen wurde oder zurückgehalten wird bzw. gesperrt ist. Wurde die Nachricht gesperrt, so wird ein eigenes Element namens <StatusWillChange> eingeschoben, welches dann die Subelemente <FutureStatus>, mit dem Wert „Usable“, und <DateAndTime>, das den Zeitpunkt angibt an dem der Status „Usable“ eintritt, enthält.

Das Element <Urgency> kann mit Werten von 1 bis 9 versehen werden, wobei 1 oberste Priorität hat. Der Wert 5 ist somit im normalen Bereich. In <Instruction> wird angegeben ob diese Nachricht fortgesetzt wird oder nicht. Mögliche Werte sind z.B. „fortsetzung“, „fortsetzung-moeglich“ oder „schluss“.[27] Wie bereits oben erwähnt kann mit dem Element <AssociatedWith> eine Verbindung zu einem anderen NewsItem ausgedrückt werden. Hier wird dem Attribut NewsItem der PublicIdentifier der assoziierten Nachricht angegeben.

[27] vgl. „NewsML by APA: Technische Dokumentation“ von der APA-IT, online unter <http://www.apa-it.at/NewsML/Deutsch/NewsMLbyAPA-TechDokumentation_v1.06.pdf> (11.06.2006)

```
<NewsManagement>
   <NewsItemType FormalName="News"/>
   <FirstCreated>20060604T035642+0200</FirstCreated>
   <ThisRevisionCreated>20060604T035642+0200
   </ThisRevisionCreated>
   <Status FormalName="Usable"/>
   <Urgency FormalName="5"/>
   <Instruction FormalName="fortsetzung-moeglich"/>
   <AssociatedWith FormalName="APASieheVorher"
   NewsItem="urn:newsml:apa.at:20060603:20060603APA0170:1"/>
</NewsManagement>
```

5.1.4 NewsComponent

Das Element <NewsComponent> enthält Metadaten die sich direkt auf die Nachricht beziehen. Hier werden die „Rollen“ der Komponenten einer Nachricht beschrieben (<Role>), die <NewsLines>, als auch <AdministrativeMetadata>, <DescriptiveMetadata> und das <ContentItem>.

Dem Element <Role> wird in der Verwendung von NewsML große Bedeutung zugeschrieben, da es mit NewsML möglich ist mehrere Komponenten in einem Dokument zu übermitteln. Diese Unterelemente können nur im <NewsComponent> angegeben werden. Durch das <Role> Element sind diese NewsComponents eindeutig zu klassifizieren, etwa als „Hauptinhalt“, „Zusatzinformation“, „FullResBild“, „ThumbNailBild“ oder „CaptionText“.[28] Die letzten drei Rollen werden für die Übertragung von Fotos verwendet. Die übrigen Elemente werden nachfolgend genauer beschrieben, da diese sich durch ihre Subelemente auszeichnen.

```
<NewsComponent>
   <Role FormalName="Hauptinhalt"/>
   <NewsLines>...</NewsLines>
   <AdministrativeMetadata>...</AdministrativeMetadata>
   <DescriptiveMetadata>...</DescriptiveMetadata>
   <ContentItem>...</ContentItem>
</NewsComponent>
```

[28] vgl. „NewsML by APA: Technische Dokumentation“ von der APA-IT, online unter <http://www.apa-it.at/NewsML/Deutsch/NewsMLbyAPA-TechDokumentation_v1.06.pdf> (11.06.2006)

5.1.5 NewsLines

In den <NewsLines> befinden sich für Menschen lesbare Text-Elemente. Hier werden der Titel (<HeadLine>), der übrigens den Titel der Meldung enthält und somit auch noch mal im <ContentItem> vorkommt, der Untertitel (<SubHeadLine>, optional verwendbar), das Copyright (<CopyrightLine>) und die Stichwörter (<KeywordLine>, optional verwendbar) gelistet. Ein Element das ebenfalls optional verwendet werden kann ist das Element <NewsLine> mit seinen Subelementen <NewsLineType> und <NewsLineText>.

Als Nachrichtenanbieter für ganz Österreich, wird in der <CopyrightLine> immer die APA genannt. Die <KeywordLine> enthält Stichwörter, welche die APA dazu verwendet um Nachrichten zu kategorisieren. Diese werden nochmals in der <DescriptiveMetadata> Anwendung finden. Wurde die Meldung aktualisiert kommt das Element <NewsLine> zum Einsatz. Das Subelement <NewsLineType> enthält mit dem Wert „UpdateEdInfo“ die Information, dass es sich um eine aktualisierte Version handelt. Im <NewsLineText> wird dann die erfolgte Korrektur angegeben.

```
<NewsLines>
    <HeadLine>China: Militärflugzeug mit 40 Insassen abgestürzt -
    Agentur</HeadLine>
    <SubHeadLine>Utl.: Im Osten des Landes - Beamter: Fünf
    Leichen geborgen </SubHeadLine>
    <CopyrightLine>Copyright, Austria Presse Agentur, Wien,
    Österreich, Alle Rechte vorbehalten</CopyrightLine>
    <KeywordLine>Unfälle/Verkehr/Luftfahrt/China</KeywordLine>
</NewsLines>
```

5.1.6 AdministrativeMetadata

Das Element <AdministrativeMetadata> enthält <Provider>, <Creator> und <Source>. Optional kann auch das Element <Contributor> verwendet werden, das jene Personen listet, die im weiteren Sinne an der Erstellung des Inhalts beteiligt waren. Als <Provider> gilt immer die APA, da sie ja Nachrichtenanbieter des KURIER Online ist. Der <Creator> Tag enthält die

Namenskürzel der Verfasser des Artikels, in dem auch mehrere <Party> Elemente vorkommen können. Die Quellen werden im <Source> Tag genannt. Hier können ebenfalls mehrere <Party> Elemente gelistet werden.

```
<AdministrativeMetadata>
    <Provider><Party FormalName="APA"/></Provider>
    <Creator><Party FormalName="vos"/></Creator>
    <Source>
       <Party FormalName="APA"/>
       <Party FormalName="Reuters"/>
       <Party FormalName="AP"/>
    </Source>
</AdministrativeMetadata>
```

5.1.7 DescriptiveMetadata

In der <DescriptiveMetadata> wird, wie der Name schon sagt, jegliche Information bereitgestellt, die den Inhalt der Nachricht beschreibt. Diese Informationen werden in die Ausdrücke <Language>, <SubjectCode>, <Location> und <Property> eingeteilt. Im <Language> Element wird die verwendete Sprache angegeben. Dies muss nach einem vorgegebenen Schema erfolgen, welches im Attribut Vocabulary referenziert wird. Der <SubjectCode> dient dazu die Nachricht in bestimmte Kategorien einteilen zu können. Im Feed der APA sind gleich zwei Kategorisierungen vorhanden.

Das APA-interne System versieht die Nachricht mit Kategorienkürzel und verweist auf die Kategoriendefinition mit dem Attribut Vocabulary. Weiters gibt es noch die Standardvariante der IPTC, welche eine achtstellige ID zur Identifikation einer Kategorie nutzt. In diesem Beispiel ist laut den IPTC Newscodes die Nachricht den Kategorien „Vermischtes“ und „Politik“ zugeordnet.[29] Die Kategorien der IPTC sind jedoch nicht auf den Tag <Subject> beschränkt, sondern können auch als <SubjectMatter> oder <SubjectDetail> angegeben werden. Der Unterschied liegt darin, dass das Element <Subject> sich nur der ersten zwei Stellen der achtstelligen Nummer

[29] vgl. „The IPTC News Codes“ von der IPTC, online unter <http://www.iptc.org/NewsCodes/index.php> (12.06.2006)

bedient und somit die Hauptkategorien anspricht. <SubjectMatter> verwendet auch die nachfolgenden drei Stellen um die Subkategorien der in den ersten zwei Stellen angegebenen Hauptkategorie zu verwenden. <SubjectDetail> geht noch eine Ebene tiefer und verwendet zusätzlich auch die letzten drei Stellen.

In <Location> wird der Ausgangspunkt der Nachricht angeführt. Im darunter gelegenen Element <Property> wird mit dem Attribut FormalName der geografisch betroffene Bereich mit den möglichen Werten „City“, „Country“, „CountryArea“, „Sublocation“ oder „WordArea“ angegeben. In den <Property> Elementen werden, wie in Kapitel 5.1.5 NewsLines erwähnt, die Keywords nochmals gelistet.

```
<DescriptiveMetadata>
    <Language FormalName="de-AT"
    Vocabulary="http://www.ietf.org/rfc/rfc3066.txt"/>
    <SubjectCode>
       <Subject FormalName="CA"
       Vocabulary="urn:newsml:apa.at:20040101:APAcat:1"/>
       <Subject FormalName="AA"
       Vocabulary="urn:newsml:apa.at:20040101:APAcat:1"/>
       <Subject FormalName="08000000"/>
       <Subject FormalName="11000000"/>
    </SubjectCode>
    <Location HowPresent="Origin">
          <Property FormalName="City" Value="Peking"/>
    </Location>
    <Property FormalName="APAstwort" Value="Unfälle"/>
    <Property FormalName="APAstwort" Value="Verkehr"/>
    <Property FormalName="APAstwort" Value="Luftfahrt"/>
    <Property FormalName="APAstwort" Value="China"/>
</DescriptiveMetadata>
```

5.1.8 ContentItem

Im letzten Element <ContentItem> ist nun der Nachrichteninhalt enthalten. Dieser kann direkt eingebettet sein, in einer externen Datei vorliegen oder in einem anderen NewsItem enthalten sein. Davor gibt es noch ein paar weitere Informationen. So wird der <MediaType> der Nachricht angegeben, bei dem es sich um „Text“, „Video“, „Audio“, „Graphic“, „Photo“ oder „ComplexData“ handeln

kann. Je nach Medientyp wird auch das zugehörige <Format> angegeben, z.B. „CSV“, „BMP“, „FLA“, „MPEG“ oder „PDF“. Dazu gibt es aber noch viele weitere mögliche Dateiformate, die angegeben werden können.[30] Im <DataContent> wird der Inhalt der Nachricht eingebettet, je nach angegebenem Format und Medientyps.

```
<ContentItem>
    <MediaType FormalName="Text"/>
    <Format FormalName="XHTML"/>
    <DataContent>
        <html xmlns="http://www.w3.org/1999/xhtml">
        <head>
            <title>China: Militärflugzeug mit 40 Insassen
            abgestürzt - Agentur</title>
        </head>
        <body>
            <p>In China ist einer Nachrichtenagentur zufolge am
            Samstag ein Transportflugzeug des Militärs mit
            mindestens 40 Menschen an Bord abgestürzt. Experten
            aus Peking leiteten den Bergungseinsatz in der
            Provinz Anhui, meldete die Agentur Xinhua am Sonntag
            unter Berufung auf Militärkreise. Fünf Leichen seien
            an der Absturzstelle im Bezirk Guangde bereits
            geborgen worden, sagte ein Beamter der
            Regionalbehörde. Guangde liegt 200 Kilometer
            südwestlich von Shanghai.</p>
         </body>
         </html>
    </DataContent>
</ContentItem>
```

5.2 Vererbung von Metadaten

Da ein NewsItem aus mehreren NewsComponents bestehen und auch auf externe NewsItems verweisen kann soll hier auch die Möglichkeit der Vererbung der Metadaten berücksichtigt werden. Hierzu gibt es laut APA einige Grundregeln:

[30] vgl. „The IPTC News Codes“ von der IPTC, online unter <http://www.iptc.org/NewsCodes/index.php> (12.06.2006)

- ***Regel:*** *Unter-NewsComponents erben grundsätzlich alle „…Metadata“-Elemente (und Sub-Elemente) der übergeordneten NewsComponent.*
- ***Ausnahme:*** *Explizit angegebene Metadaten in Unter-NewsComponents überschreiben allenfalls geerbte Metadaten.*
- ***Grenzen der Regel:*** *Metadaten können nicht an NewsItems in Unter-NewsComponents vererbt werden. (Das trifft bei „NewsML by APA“ immer dann zu, wenn eine Unter-NewsComponent auf eine extern vorliegende NewsML-Instanz verweist.) (Zitat)* [31]

So verfügt also jedes NewsItem über die in ihm angegebenen Metadaten und kann sie nicht an andere NewsItems weitervererben. Die darunter liegenden NewsComponents und deren Unter-NewsComponents bedienen sich der Metadaten aus dem direkt über ihnen gelegenen NewsItem. Die NewsComponents müssen aber nicht die Metadaten des NewsItems darüber erben. Sie können auch eigene Metadaten enthalten, welche dann die vererbten Informationen überschreiben.

[31] „NewsML by APA: Technische Dokumentation“ von der APA-IT, online unter <http://www.apa-it.at/NewsML/Deutsch/NewsMLbyAPA-TechDokumentation_v1.06.pdf> (11.06.2006)

Hierzu soll wiederum eine Grafik zur Veranschaulichung und leichterem Verständnis dienen.

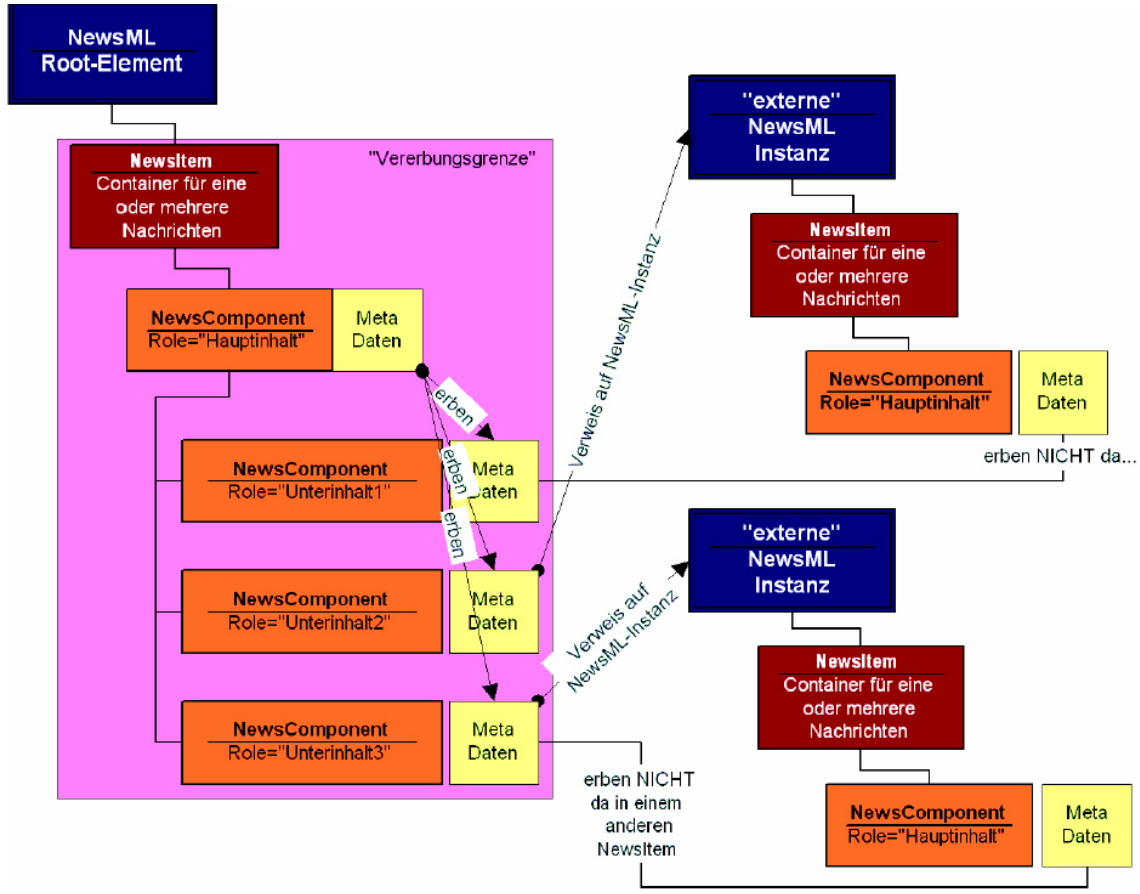

Abbildung 3: Darstellung von Vererbung der Metadaten

6. Integration von „NewsML by APA“

Der APA-OnlineManager (AOM) stellt für die Redakteure des KURIER Online die wichtigste Nachrichtenquelle dar. Dieses Tool wird, wie der Name schon vermuten lässt, direkt von der APA zur Verfügung gestellt und gibt Auskunft über aktuelle Geschehnisse. So können die Redakteure nach bestimmten Artikeln suchen oder sich einfach die Tagesnachrichten ansehen, um dann aus den vorhandenen Stories auszuwählen. Dabei gibt die APA neben der eigentlichen Nachricht auch Metainformationen, wie z.B. Quellen, APA-interne Ressorteinteilungen oder den Erscheinungszeitpunkt einer Nachricht an.

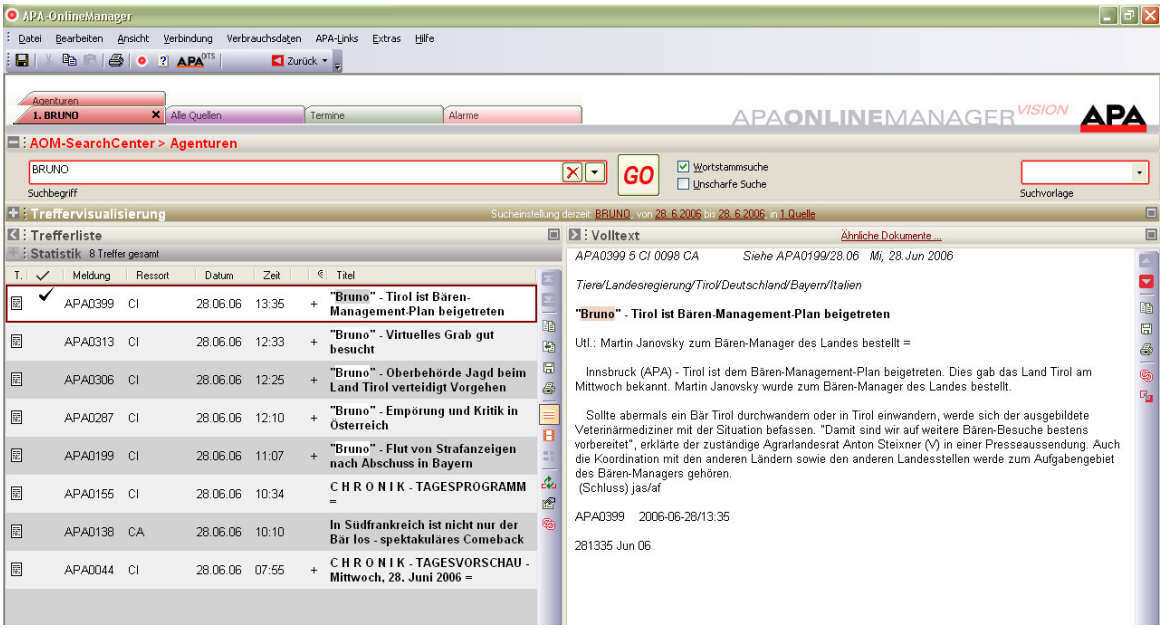

Abbildung 4: APA-OnlineManager (Screenshot)

Die APA stellt aber nicht nur Anwendertools wie den AOM zur Verfügung, sondern bietet ihren Nachrichtenabnehmern unter anderem auch die Möglichkeit einer Datenlieferung in Form von NewsML. Der APA-Basisdienst im NewsML-Format wird von KURIER Online bereits seit Februar 2006 automatisiert per FTP (File Transfer Protocol) bezogen. Seitens des KURIER Online werden diese

NewsML-Feeds aber noch nicht effektiv genutzt. Das Problem hieran ist die fehlende Schnittstelle zum CMS.

KURIER Online verwendet ein CMS namens „Digital Workroom“ (kurz „DW“ genannt, online unter http://www.digitalworkroom.com/), welches sich einfach erweitern und steuern lässt. So hat man die Möglichkeit, das CMS den Wünschen anzupassen und stößt nicht auf unüberwindbare Grenzen in der Benutzung. Durch die gegebene Möglichkeit der einfachen Erweiterung, bietet es sich an die Schnittstelle „NewsML zu CMS“ selbst umzusetzen.

6.1 Einführung in das KURIER-CMS

Der DW ist an der Benutzeroberfläche sehr strukturiert aufgebaut. Als Nutzer (Redakteur) muss man zu allererst die gewünschte Plattform „kurier“ wählen um dann eine Auswahl der vorhandenen Ressorts zu bekommen. Ist ein Ressort ausgewählt z.B. „Nachrichten | Österreich“, hat man verschiedene Möglichkeiten der weiteren Vorgehensweise. Man kann entweder in einer leeren Maske einen neuen Artikel für dieses Ressort anlegen (mittlerer Frame), oder links nach bereits vorhandenen Artikeln suchen.

Hat man einen Artikel in der Maske erstellt bzw. bearbeitet wird dieser mittels des Buttons „Erstellen“ bzw. „Aktualisieren“ in die dahinter liegende Datenbank geschrieben. Wird der Button „Export“ gedrückt wird der Artikel statisch am Webserver abgelegt und erscheint online.

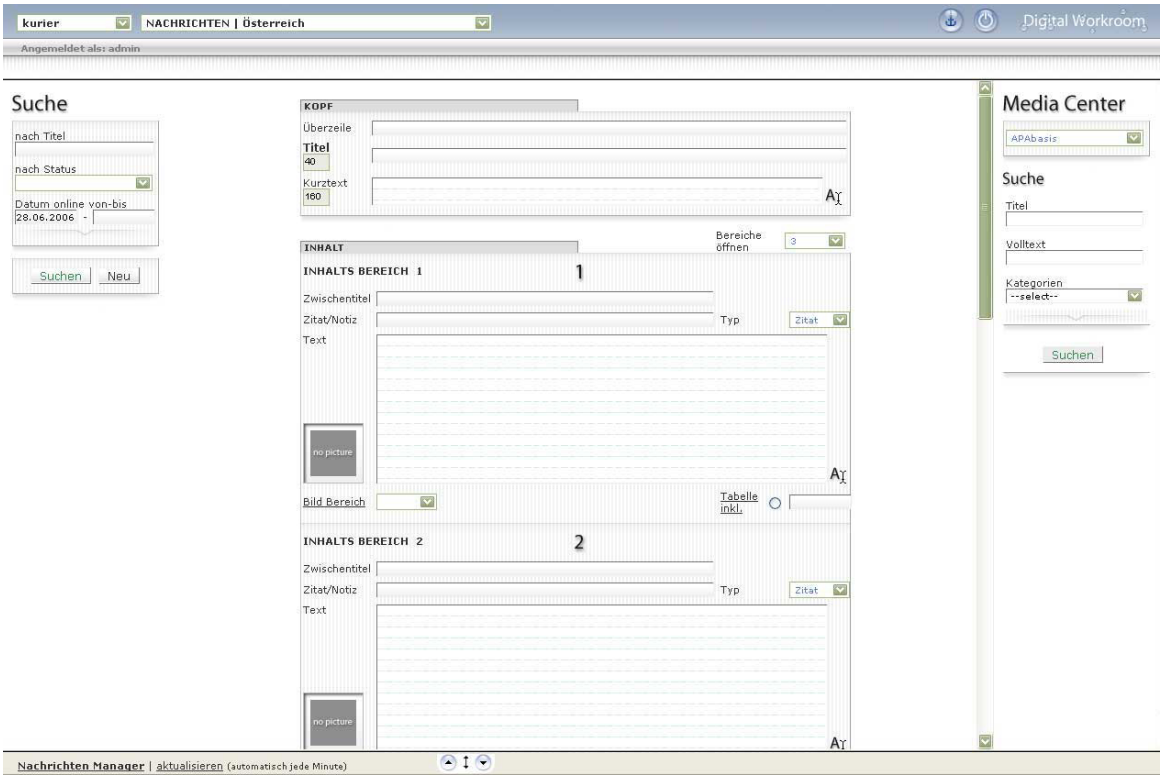

Abbildung 5: Das KURIER Online CMS „Digital Workroom“ (Screenshot)

Damit ein Artikel das gewisse Etwas bekommt, kann man z.B. Bilder, ganze Bildergalerien, Dokumente und seit neuestem auch Videos mit dem Artikel verknüpfen. Die Einbindung dieser Features wird über das „Media Center“ im rechten Bereich gesteuert. Hier können verschiedene Dateiformate upgeloadet werden. Weiters ist es möglich nach bereits vorhandenen Dateien zu suchen um z.B. bereits erstellte Bildergalerien zu überarbeiten und wieder verwenden zu können. Das „Media Center“ bietet also für die Einbindung des NewsML-Feeds genau den richtigen Rahmen.

Aus technischer Sicht wird das CMS im Hintergrund über so genannte Layouts gesteuert, die mittels einer „Template Engine“ namens „Netautor“ verwaltet werden. Diese Templates bilden die absolute Grundlage für die Funktionalität und auch das Erscheinungsbild des DW’s. Ein Layout wird meistens in mindestens zwei verschiedenen Programmiersprachen geschrieben. Es wird in einer netautoreigenen Syntax verfasst, aber meist noch zusätzlich durch PHP erweitert (oder auch umgekehrt). Der Netautor ist auch für die „Verwaltung“ der dahinter

liegenden Datenbank zuständig. Mit Hilfe dieses Tools werden die Rollen der einzelnen Tabellen klar strukturiert. So werden beispielsweise in der „content-table“ nur Inhalte verwaltet die Content für kurier.at darstellen oder der strukturelle Aufbau von kurier.at in der „structure-table“ festgehalten. Diese sind aber natürlich nicht die einzigen Tabellen die diesem Schema unterliegen, es gibt noch eine Fülle an weiteren Tabellen.

6.2 Einbettung von NewsML in das vorhandene CMS

Wie bereits erwähnt ist das CMS von KURIER Online relativ einfach zu erweitern. Um ein Script ausführen zu können, wird ein neues Layout erstellt, welches dann im DW noch referenziert werden muss. Dies wird im Administrationsbereich festgelegt (siehe Abbildung 6). Hier wird die neue Option „APAbasis“ ins „Media Center“ aufgenommen und mit der beim Aufruf auszuführenden Layoutnummer versehen.

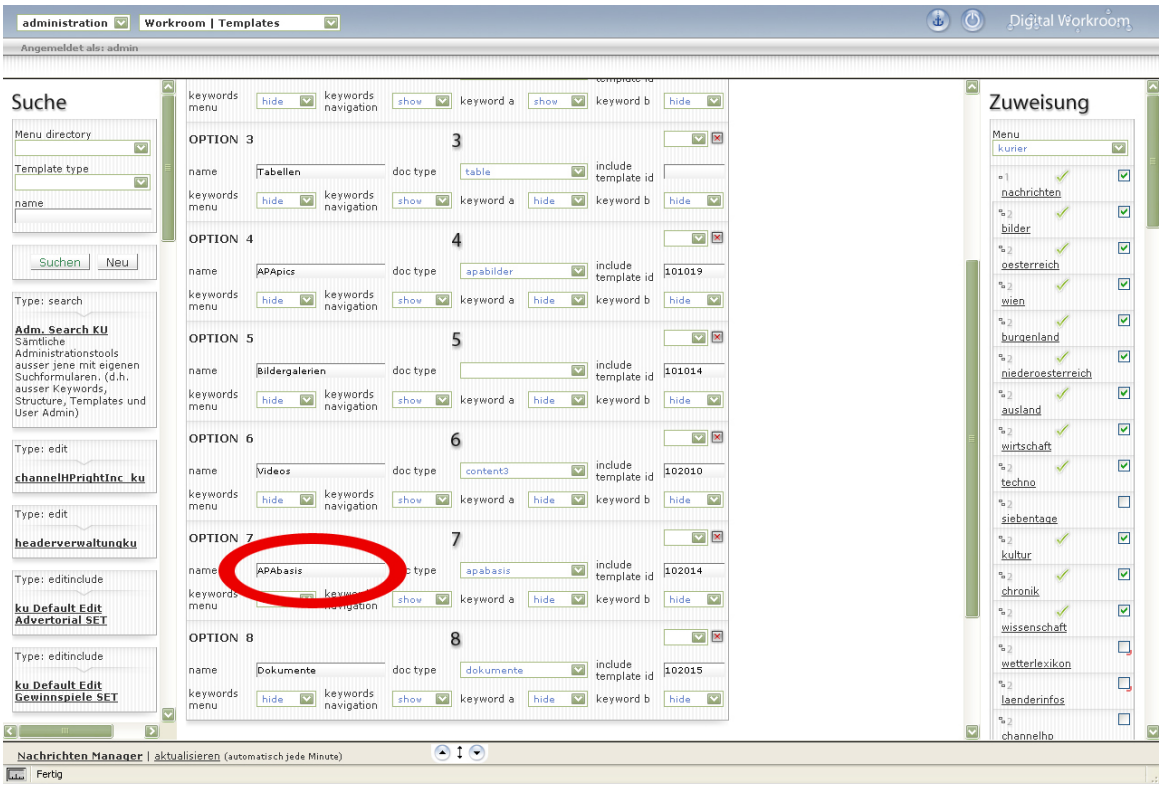

Abbildung 6: Hinzufügen eines neuen Layouts (Screenshot)

Ist dies geschehen, kann die neue Option über das „Media Center“ verwendet werden. Der Inhalt der gewählten Option kommt dabei, wie bereits erwähnt, direkt aus dem aufgerufenem Template.

Da die beiden zu verbindenden Elemente, NewsML auf der einen Seite und das CMS samt dahinter liegender Datenbank auf der anderen Seite, auf unterschiedlichen Servern angesiedelt sind, musste für die Darstellung im DW ein „Work-Around“ implementiert werden. Um eine gute und schnelle Verarbeitung zu erzielen, ohne Einfluss auf die angesiedelte Serverumgebung nehmen zu müssen, wurde das Script zur Verwaltung der Dokumente auf demselben Server angesiedelt, auf dem sich auch die NewsML-Dateien befinden. So wird das Durchsuchen der Dokumente aus dem Uploadordner erleichtert. Durch den Aufruf eines externen Scripts werden die Suchergebnisse im CMS eingebunden und die serverübergreifende Interaktion durch die Abfrage eines speziell generierten Ausdrucks abgesichert.

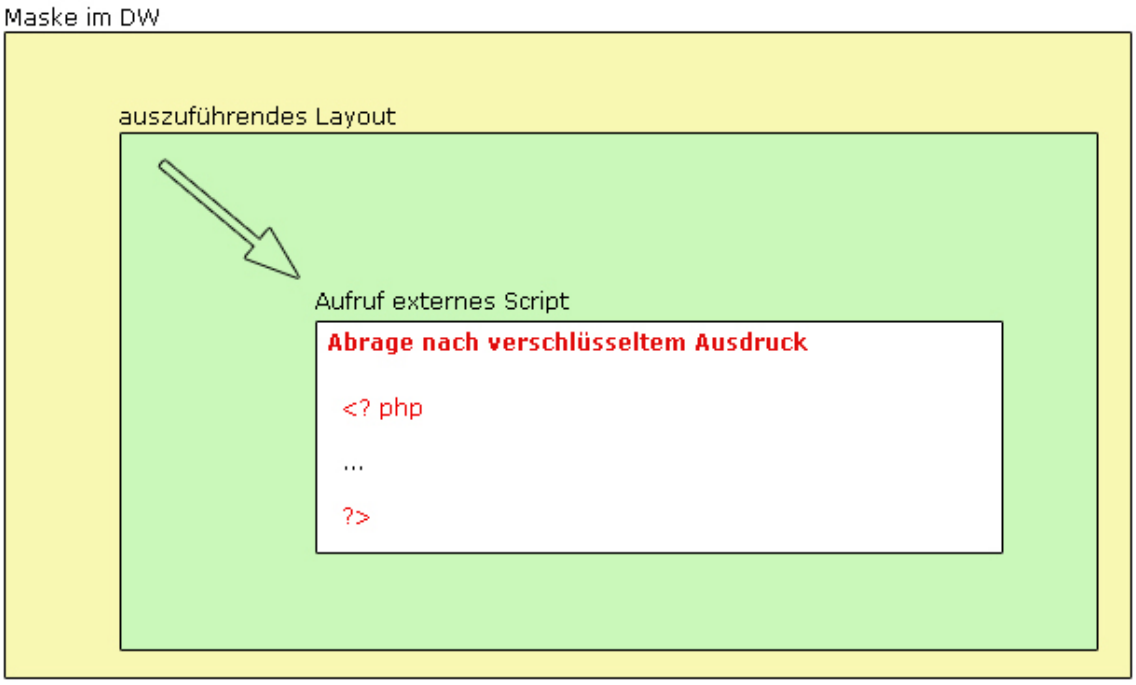

Abbildung 7: Grafische Darstellung der Einbindung des Scripts ins CMS

Die vorhandenen NewsML-Dateien werden also nur dann im „Media Center“ angezeigt, wenn sie einer eingegebenen Suchabfrage entsprechen. So wird vermieden, dass die Redakteure durch das Laden aller verfügbaren NewsML-Dokumente einer relativ langen Ladezeit ausgeliefert sind und unnötig RAM-Speicher vom Server belegt wird. Immerhin befinden sich im Uploadordner ca. 2000 Dokumente, was die Durchlaufzeit von Abfragen recht deutlich erhöht. Die Redakteure müssen also mindestens eine Suchoption angeben um einen Artikel zu finden. Hierbei können sie die Kategorie durchsuchen oder nach einem Stichwort im Titel oder im Text suchen. Es werden jedoch maximal 100 Treffer angezeigt. War die Suche erfolgreich und ein Stichwort wurde im Titel der Nachricht gefunden, wird dieses mittels einer Farbhinterlegung hervorgehoben. Es werden aber noch weitere Hervorhebungen durch den Einsatz von Farbe bewerkstelligt.

Da die Nachrichten mit unterschiedlichen Dringlichkeitseinstufungen ausgezeichnet sind, werden alle Nachrichten die über der normalen Priorität (5) liegen farblich markiert. So werden sehr wichtige Nachrichten (1) mit roter Schrift, wichtige Nachrichten (2) mit oranger Schrift, die nächsten Stufen 3 und 4 mit blau und grün dargestellt. Die Redakteure sollen dadurch auf diese Meldungen aufmerksam gemacht werden um diese vorrangig zu behandeln und so schnell als möglich online zu stellen.

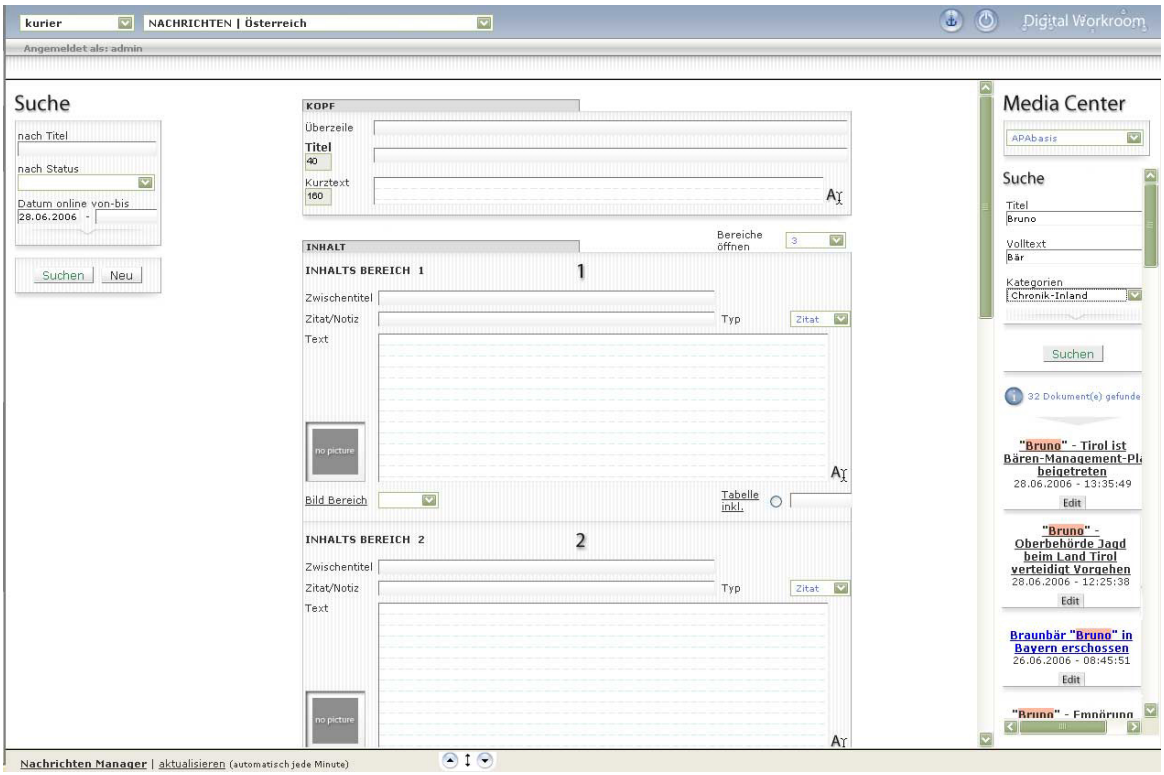

Abbildung 8: Artikelsuche nach Stichwörtern und Kategorie (Screenshot)

Die nach den eingestellten Suchoptionen gefundenen Nachrichten werden vorerst nur mit dem Titel der Nachricht angezeigt. So ist die Überprüfung des Inhalts natürlich noch nicht gegeben. Durch Klicken auf den Titel öffnet sich ein Popup-Fenster, in dem nun alle relevanten Informationen dargestellt werden. Neben dem Titel, dem möglichen Untertitel, Ort, Quelle und dem Text sind noch Angaben zum Verfasser der Nachricht, der Erstellungszeitpunkt und die Kategorien nach APA-Zuteilung vorhanden. Zusätzlich wird, wenn angegeben, auch auf die assoziierte Nachricht verlinkt (siehe Kapitel 5.1.3 NewsManagement).

Abbildung 9: Ansicht eines APA-Artikels im Detail (Screenshot)

Um das Zusammenspiel der einzelnen Elemente zu verstehen, wird hier vorab die Serverlandschaft mit ihren Zusammenhängen dargestellt.

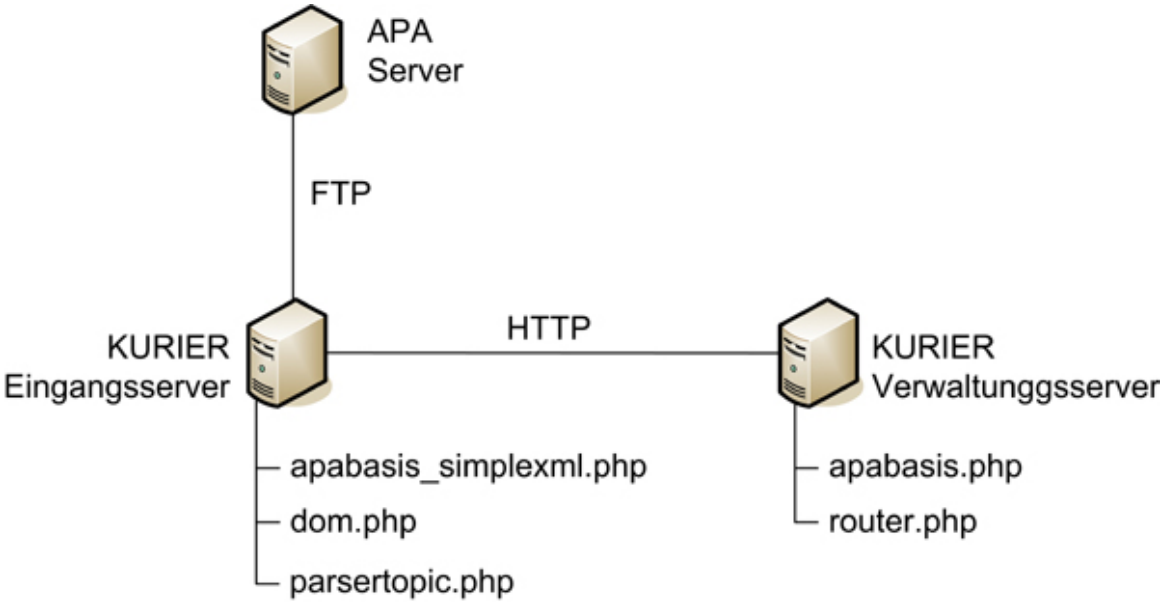

Abbildung 10: Skizze der Serverlandschaft

Weiters muss erwähnt werden, dass sich alle Pfade aus Sicherheitsgründen auf die nötigsten Informationen beschränken und durch die Textmarke {Pfad} ersetzt wurden.

6.2.1 Die Theorie

Zu Beginn wurden alle potenziellen Dateien (der Status musste „Usable“ sein) in ein Array gespeichert und für den Gebrauch im CMS zur Verfügung gestellt. Diese Umsetzung stellte sich jedoch als zu belastend für den Server heraus, da die Inhalte nicht aus einer Datenbank kamen, sondern direkt aus dem NewsML-Dokument ausgelesen wurden. So hat die Suche, die jeden Eintrag des Arrays durchsuchen musste, dementsprechend lange gedauert. Aus diesem Umstand heraus entstand die Idee nicht den gesamten Inhalt der NewsML-Dateien zu übernehmen, sondern nur jene Inhalte, die für die Darstellung relevant waren.

Als nächstes war das Auslesen der NewsML-Dokumente per XSLT (Extensible Stylesheet Language Transformation) und XPath (Kurzform von XML Path Language) geplant sowie die gewonnenen Inhalte in eine komprimiertere Form zu bringen. Hier das Code-Snippet:

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
version="1.0">
<xsl:template match="/">
<bezML>
<field name="xmlval_TITLE"><xsl:value-of
select="(//HeadLine)"/></field>
<field name="xmlval_TITLE"><xsl:value-of
select="(//SubheadLine)"/></field>
</bezML>
</xsl:template>
</xsl:stylesheet>
```

Nach einigen gescheiterten Versuchen dann die Gewissheit: Die Server des KURIER Online haben keine XSL-Erweiterung installiert. Somit musste eine neue Möglichkeit ins Auge gefasst werden.

Die Alternative sollte AJAX (Asynchronous JavaScript and XML) heißen. Das funktionierte in der Testumgebung mit einer geringen Anzahl an NewsML-Dokumenten sehr gut. Hier der Aufruf der Schnittstelle:

```
<script type="text/javascript">
try{
   //firefox, ie 7.0,...
   var xmlhttp = new XMLHttpRequest();
}catch(e){
   //ie 6.0
   var xmlhttp = new ActiveXObject("Microsoft.XMLHTTP");
}
</script>
```

Der Aufruf der Funktion `ajax()` übergibt den Dateinamen, um vom gewünschten File Daten zu holen und den Inhalt in die Formularfelder zu füllen:

```
<script type="text/javascript">
function ajax(file){
     xmlhttp.open("GET","{Pfad}/dom.php?file="+file+".xml",true);
     xmlhttp.onreadystatechange=function(){
          if(xmlhttp.readyState==4){
```

```
                alert(xmlhttp.responseText);

                var headline = xmlhttp.responseXML.getElementsBy
                TagName("title")[0].childNodes[0].data;
                var summary = xmlhttp.responseXML.
                getElementsByTagName("summ")[0].childNodes[0].
                data;

                var content = xmlhttp.responseXML.
                getElementsByTagName("body")[0].childNodes[0].
                data;

                var credit = xmlhttp.responseXML.
                getElementsByTagName("cred")[0].childNodes[0].
                data;
            }

            parent.edit.document.form1.
            elements['xmlval_TITLE[0]'].value = headline;

            parent.edit.document.form1.
            elements['xmlval_SUMMARY[0]'].value = summary;

            parent.edit.document.form1.
            elements['xmlval_CONTENT[0][C_TEXT][0]'].value =
            content;

            parent.edit.document.form1.
            elements['xmlval_CREDIT[0]'].value = credit;
      }
      xmlhttp.send("");
}
</script>
```

Die mit DOM (Document Object Model) gewonnenen Inhalte wurden in eine neue Form gebracht und als `<bezML>` zum angeforderten File übergeben (vgl. Kapitel 6.2.4.5 dom.php):

```
<bezML>
<title>".$headline.'</title>
<summ>'.$subheadline.'</summ>
<body>'.$body.'</body>
<cred>'.$sourcen.'</cred>
</bezML>
```

Den Praxistest bestand diese Variante dann leider nicht, da der XMLHttpRequest nur im Bereich der aktuell verwendeten Domain funktioniert. Andere Domains

und auch Subdomains können nicht „requested“ werden.[32] Da aber das Durchsuchen der Files am Server des FTP-Upload Ordners und die Darstellung sowie die Übernahme am Verwaltungsserver geschehen musste, konnte hier keine einfache Umgehung des Problems realisiert werden.

Eine Lösung für dieses Problem wäre eine Weiterleitung vom Verwaltungsserver auf den Eingangsserver gewesen. Diese Art der Problemlösung wurde allerdings zu spät erkannt und da bereits eine zufriedenstellende Lösung existiert, wurde dieser Ansatz nicht weiter verfolgt. Es wäre also eine zusätzliche Möglichkeit geschaffen worden, die eine bewährte Methode ersetzt, was in diesem Fall nicht sinnbringend ist.

Eine andere Variante wäre der direkte Datenbankeintrag gewesen. Dies erschien aber bei näherer Betrachtung als zu umständlich, da der unbearbeitete Artikel übernommen werden würde, der Redakteur ihn dann über die Suche im CMS aufrufen müsste und dann erst bearbeiten könnte – eine suboptimale Lösung.

6.2.2 Die Performance

Die Performancetests wurden über die PHP-Funktion `microtime()` durchgeführt, die die Laufzeit des Scripts in Mikrosekunden zurückgibt. Man kann sagen, dass sich die Performance nach und nach mit der Ausreifung des Prozesses verbesserte. Zu Beginn, mit der Übergabe des gesamten Arrays, dauerte die Ausführung und Anzeige noch relativ lange. Ein Stichprobentest ergab drei Sekunden bei einer Anzahl von 1278 Dokumenten, was sich gut anhört. Wenn man jedoch damit arbeitet sind drei Sekunden sehr lang. Bei einer höheren Anzahl an Dokumenten stieg natürlich auch die Durchlaufzeit an. So wurden bei einer Anzahl von 2573 Dokumenten eine Durchlaufzeit von zwölf Sekunden gemessen, was absolut nicht zumutbar ist. Weiters wurde durch die riesigen Mengen an Daten auch unnötig der RAM-Speicher des Servers belegt, was die Nutzung stark beeinträchtigte.

[32] vgl. „Mein XMLHttpRequest und ich“ von DenkZEIT, online unter <http://www.gungfu.de/facts/archives/2005/07/29/mein-xmlhttprequest-und-ich/> (21.08.2006)

Bei der neuen Lösung, welche nur jene Dokumente im Array übergibt, die den eingegebenen Suchkriterien entsprechen, läuft dieser Prozess schon wesentlich schneller ab. Hier musste die Anzeige jedoch auch auf maximal 100 Dokumente begrenzt werden, um nicht Gefahr zu laufen wieder an die 1000 Dokumente zu übergeben und den RAM-Speicher für andere Anwendungen zu blockieren. Hierzu ist anzumerken, dass es durchaus sein kann, dass der Online Redakteur nur eine Kategorie als Suchkriterium angibt. Im Laufe des Tages werden jedoch, je nach Produktivität der APA, bis zu 2500 Dateien oder mehr empfangen. Es kann also durchaus der Fall sein, dass einer Kategorie weitaus mehr als 100 Dateien zugeordnet sind, was sich wiederum auf die Durchlaufzeit auswirken würde. Nun werden also maximal die 100 aktuellsten Dokumente ins Array gespeichert, die den gewünschten Suchkriterien entsprechen. So läuft der Prozess schon wesentlich schneller ab.

Laut Messungen dauert der Aufbau der Suchmaske, also das Laden von „apabasis.php“ ohne Suchanfrage ca. 14 Nanosekunden. Mit der Kategorie als einzigem Suchkriterium werden 100 Treffer in ca. 105 Nanosekunden angezeigt. Je nach Suchwort im Titel dauert das Anzeigen der Treffer hier 500 bis 650 Nanosekunden. Für die Textsuche werden je nach Suchwort zwischen 500 Nanosekunden und einer Sekunde benötigt. Insgesamt läuft der Prozess nun viel schneller und stabiler, sodass die Suche ohne lange Wartezeiten benutzbar ist und der Server nicht belastet wird.

6.2.3 Die Fehleranfälligkeit

Die Fehleranfälligkeit des ganzen Systems muss von verschiedenen Seiten beleuchtet werden, denn ein Fehler kann immer und überall auftreten und ist nie auszuschließen.

Der Feed der APA wird weitgehendst überprüft bevor er übermittelt wird. So wird bereits bei der Erstellung des NewsML-Dokuments auf etwaige Fehler geprüft und die Probleme, wenn möglich, direkt im Script gelöst und mitprotokolliert. Durch die Protokollierung können Fehler die nicht im Script gelöst werden

konnten, später nachvollzogen und eine andere Lösung des Problems angestrebt werden. Weiters könnte auch ein Fehler in der Übermittlung der NewsML-Dokumente auftreten. Laut APA könnte sich die Zustellung der Meldungen in diesem Fall verzögern, ein Totalausfall des Erhalts dieser Dokumente sei aber ziemlich unwahrscheinlich, da auch hier Vorkehrungen getroffen wurden.

Auch seitens des KURIER Online wurden einige Fehlerfälle berücksichtigt, die scriptseitig abgefangen werden und bei Eintreten ein Informationsmail versenden. Davon sind die Scripte „parsertopic.php“, „apabasis_simplexml.php“, „dom.php“, „router.php“ und „apabasis.php“ betroffen. In der Datei „parsertopic.php“ wird der Fehlerfall der fehlenden Kategorien berücksichtigt, welche direkt von der Website der APA-IT geladen werden. Sollten diese aus irgendeinem Grund nicht geladen werden können, wird das lokal gesicherte Topicset der APA verwendet. In „apabasis_simplexml.php“ wird beim Auslesen der Dateien aus dem Ordner unter anderem nach der Dateigröße abgefragt, die nicht null sein darf. Zusätzlich wird nach dem Auslesen das Array überprüft, in dem die Dateinamen gespeichert wurden sowie das Laden der Dateien bevor sie nach den übergebenen Kriterien durchsucht werden. In der Datei „dom.php“ wird vor dem Laden durch DOM geprüft, ob die angeforderte Datei existiert. In „router.php“ wird abgefragt, ob die Inhalte aus „dom.php“ übernommen werden. Im Script „apabasis.php“ wird überprüft ob die APA-Kategorien richtig übernommen wurden. Ist dies nicht der Fall wird das Drop-Down Menü nicht angezeigt. Ebenso wird hier abgefragt, ob das Array mit den Suchergebnissen aus der Datei „apabasis_simplexml.php“ ausgelesen werden konnte.

6.2.4 Die Implementierung

Für die Darstellung der Inhalte im CMS werden letztendlich sechs Dokumente benötigt: „functions.inc.php“, „apabasis_simplexml.php“, „dom.php“, „parsertopic.php“, „apabasis.php“ und „router.php“, wobei hier nach der physischen Ansiedlung der Dokumente differenziert werden muss. Die Dateien „apabasis_simplexml.php“, „dom.php“ und „parsertopic.php“ befinden sich am Eingangsserver, „apabasis.php“ und „router.php“ befinden sich am

Verwaltungsserver. Die Ausnahme bildet die Datei „functions.inc.php“, welche sowohl am Eingangsserver als auch am Verwaltungsserver vorhanden ist.

6.2.4.1 Die Serverumgebung

Die Umsetzung der Schnittstelle „NewsML zu CMS“ passiert auf zwei Servern, dem Eingangsserver und dem Verwaltungsserver. Am Eingangsserver werden die NewsML Dokumente, die per FTP von der APA geschickt werden, verwaltet. Hier ist die PHP Version 5.1.2 verfügbar und unter anderem die DOM-Erweitung sowie die SimpleXML-Erweiterung installiert. Am Verwaltungsserver ist das CMS angelegt welches die Inhalte von kurier.at verwaltet. Hier ist die PHP Version 4.3.9 verfügbar und weder DOM noch SimpleXML implementiert.

6.2.4.2 Das Layout

Das zuvor im CMS definierte Layout 102014 (siehe Abbildung 6) beinhaltet den Aufruf der Datei „apabasis.php“ und sorgt so dafür, dass innerhalb der CMS-Umgebung dieses eigentlich externe Script ausgeführt wird.

```
<? include("{Pfad}/apabasis.php"); ?>
```

6.2.4.3 functions.inc.php

In dieser Datei werden zentral alle selbst geschriebenen Funktionen verwaltet. Es enthält die vier Funktionen `simplexml2array()`, `highlight()`, `textfarbe()` und `repl()`. Die Funktion `simplexml2array()` wandelt SimpleXML-Elemente in ein Array um, `highlight()` markiert im Titel gefundene Suchwörter, `textfarbe()` markiert die Dringlichkeit der Mitteilung und `repl()` ersetzt bestimmte Zeichen.

```
<?
function simplexml2array($xml) {
     if (get_class($xml) == 'SimpleXMLElement') {
          $attributes = $xml->attributes();
          foreach($attributes as $k=>$v) {
               if ($v) $a[$k] = (string) $v;
```

```
            }
            $x = $xml;
            $xml = get_object_vars($xml);
      }
      if (is_array($xml)) {
            if (count($xml) == 0) return (string) $x;
            foreach($xml as $key=>$value) {
                  $r[$key] = simplexml2array($value);
            }
            return $r;
      }

      return (string) $xml;
}

function highlight($pattern,$inhalt) {
      $suche = preg_replace("#(".$pattern.")#i","<span
      style=\"background-color:#ffbaa2\">\${1}</span>",$inhalt, -
      1);
      return $suche;
}

function textfarbe($urgency) {
      switch ($urgency) {
            case 1: $color="#CC0000"; //rot
                        break;
            case 2: $color="#EE9900"; //orange
                        break;
            case 3: $color="#0000DD"; //blau
                        break;
            case 4: $color="#00AA00"; //grün
                        break;
            default: $color="#333333"; //dunkelgrau
      }
      return $color;
}
function repl($var) {
      $var = strip_tags(strtr($var, array("<br>"=>" ",
      "'"=>"&#39;", "\""=>"&#34;", "Utl.: "=>"", "Utl.:"=>"",
      "\n"=>" ")));
      return $var;
}
?>
```

Die Funktion `repl()` sorgt dafür, dass Zeichen die nicht richtig ausgewertet werden bzw. nicht dargestellt werden sollen, im Vorfeld ersetzt werden. Durch die verwendete PHP-Funktion `htmlspecialchars()` werden die Zeichen &, <, >, “ und ’ in HTML-Code (z.B. & für das &-Zeichen) bzw. Unicode (' für das einfache Anführungszeichen) dargestellt. Die PHP-Funktion `strip_tags()`

entfernt alle HTML- und PHP-Tags und schützt so vor der Einschleusung von ausführbarem Code.

6.2.4.4 apabasis_simplexml.php

Die Datei „apabasis_simplexml.php“ beschreibt die Einbettung von „NewsML by APA“ durch Anwendung der SimpleXML-Funktionen, die PHP 5 zur Verfügung stellt. Dieses Script ist am Eingangsserver des KURIER Online angesiedelt und wird aus der Datei „apabasis.php“ aufgerufen.
Zuerst wird nach dem übergebenen Ausdruck abgefragt, fehlt dieser, wird das Script abgebrochen. Beim Auslesen der Dateien aus dem Ordner werden diese nach verschiedenen Kriterien ausgesucht. So darf kein Ordner und nur Dateien mit dem Suffix „xml“ und einer Dateigröße über null Bytes ausgelesen werden. Danach werden die ausgelesenen Dateinamen des NewsML-Basisdienstes (nur Text) im Array `$file_arr` mit dem Zeitpunkt der Dateiänderung als Schlüssel gespeichert. Damit ist es möglich die im Array befindlichen Dateinamen nach ihrem Eingangszeitpunkt zu sortieren.

```
<?
$pfad = "/apa/basis/";
$verz = opendir($pfad);

while (false !== ($file = readdir($verz))) {
      if($file != "." && $file != ".." && ereg('.xml', $file) &&
      filesize($pfad.$file)!=0) {
        $file_arr[filemtime($pfad.$file)] = $file;
      }
}
?>
```

An dieser Stelle wird das erstellte Array `$file_arr` überprüft, um sicherzugehen, dass es Einträge enthält. Ist dies nicht der Fall wird ein Mail mit einer Fehlermeldung versendet.

```
<?
$anz_files = count($file_arr);

if($anz_files==0) {
```

```
      mail("bettina.tuller@kurier.at","Fehler in
      apabasis_simplexml.php", "keine Files im Array! Bitte
      überprüf mich!");
}
?>
```

Ist das Array `$file_arr` befüllt, werden die darin gespeicherten Werte mittels der Funktion `krsort` sortiert, welche ein Array nach ihren Schlüsseln in umgekehrter Reihenfolge sortiert. Der Name des aktuellsten Dokuments steht somit als erster Wert im Array. Danach wird es mit einer `foreach`-Schleife durchlaufen. Hierbei wird versucht jedes Dokument mit der vorgefertigten PHP-Funktion `simplexml_load_file` zu laden. Misslingt dies, wird wiederum eine Fehlermeldung per Mail versendet. Ansonsten wird der Inhalt der Datei, wenn der Status des Dokuments mit „Usable“ freigegeben ist, nach den Suchanfragen, die von „apabasis.php“ als Variablen übergeben werden, durchsucht.

```
<?
krsort($file_arr);

foreach($file_arr as $key => $file) {
      if($xmlFile = simplexml_load_file($pfad.$file)) {
         $status = $xmlFile->NewsItem->NewsManagement->Status[0]
         ->attributes()->FormalName;

         if($status=="Usable") {
            {Abarbeiten der Suchanfragen}
         }
      } else {
          mail("bettina.tuller@kurier.at", "Fehler in
          apabasis_simplexml.php", "Fehler beim Versuch folgende
          Dateien zu laden:<br><br>".$inhalt);
      }
}
?>
```

Wie man sieht werden die Elemente als Objekte angesprochen (`$xmlFile->NewsItem->NewsManagement->Status[0]->attributes()->FormalName;`). SimpleXML bietet aber durch die angebotenen PHP-Funktionen auch gute Möglichkeiten zur weiteren Verarbeitung von XML. Hierbei ist es möglich die Elemente in XML-Form zu verwandeln (asXML()) oder aber eine XPath-Abfrage auszuführen (xpath()).

Für die Abarbeitung der Suchanfrage werden drei Elemente aus dem NewsML-Dokument benötigt:

```
<?
$title = $xmlFile->NewsItem->NewsComponent->ContentItem
          ->DataContent->html->head->title;

$body = $xmlFile->NewsItem->NewsComponent->ContentItem
         ->DataContent->html->body;

$subject = $xmlFile->NewsItem->NewsComponent->DescriptiveMetadata
             ->SubjectCode;
?>
```

Die per GET-Methode übermittelten Suchbegriffe werden in lokale Variablen übernommen und nacheinander abgearbeitet. Die Suchoptionen können unabhängig voneinander benutzt werden, werden jedoch je nach angegebener Option unterschiedlich behandelt. Wird eine Kategorie angegeben hat sie die höchste Priorität, da sie ja quasi der „Container“ einer Nachricht ist. Danach werden der Titel und zuletzt der Text ausgewertet. Die Variable `$counter` zählt die Anzahl der ausgeführten Speicherungen ins Array um die Performancezeit gering zu halten.

```
<?
if($counter == 100) break;
$patternText= "#(".$_GET["patterntext"].")#i";
$patternTitle= "#(".$_GET["patterntitle"].")#i";
$category= $_GET["category"];

// ist die Kategorie definiert und in Titel und Text ein
// Suchbegriff enthalten wird die erste Anweisung ausgeführt

if($category != "0") {
   if($patternTitle != "#()#i" && $patternText != "#()#i") {
      if (($subject->Subject[0]->attributes()->FormalName ==
      $category) &&
      (@preg_match($patternTitle,strtolower($title)))){
          foreach($body as $weitarr) {
              foreach($weitarr as $text) {
                  if (@preg_match($patternText,
                  strtolower($text))){
                      if(!array_key_exists($file,$such_array)) {
                         $such_array[$file]=$xmlFile;
                         $counter++;
                      }
                  }
```

```
                }
            }
        }

// sind nur Kategorie und Titel definiert wird diese Anweisung
// ausgeführt

   } elseif($patternTitle != "#()#i") {
         if (($subject->Subject[0]->attributes()->FormalName ==
         $category) && (@preg_match($patternTitle,
         strtolower($title)))) {
             if(!array_key_exists($file,$such_array)) {
                $such_array[$file]=$xmlFile;
                $counter++;
             }
         }

// sind nur Kategorie und Text definiert wird diese Anweisung
// ausgeführt

   } elseif ($patternText != "#()#i") {
         if ($subject->Subject[0]->attributes()->FormalName ==
         $category) {
             foreach($body as $weitarr) {
                 foreach($weitarr as $text) {
                     if (@preg_match($patternText,
                     strtolower($text))) {
                         if(!array_key_exists($file,$such_array))
                         {
                            $such_array[$file]=$xmlFile;
                            $counter++;
                         }
                     }
                 }
             }
         }

// sind beide, Text und Titel, leer, wird nur nach der Kategorie
// durchsucht

   } else {
         if ($subject->Subject[0]->attributes()->FormalName ==
         $category) {
             if(!array_key_exists($file,$such_array)) {
                $such_array[$file]=$xmlFile;
                $counter++;
             }
         }
   }
}

// ist keine Kategorie angegeben wird nach den Suchbegriffen in
// Text und Titel gesucht

if($patternTitle != "#()#i" && $category == "0") {
```

```
   if ($patternText != "#()#i") {
       foreach($body as $weitarr) {
           foreach($weitarr as $text) {
               if (!array_key_exists($file,$such_array) &&
               (@preg_match($patternTitle, strtolower($title)))
               && (@preg_match($patternText, strtolower($text))))
               {
                   $such_array[$file]=$xmlFile;
                   $counter++;
               }
            }
       }

// ist der Suchtext leer, wird nur nach dem Titel gesucht

   } else {
         if (@preg_match($patternTitle, strtolower($title))){
             if(!array_key_exists($file,$such_array)) {
               $such_array[$file]=$xmlFile;
               $counter++;
             }
         }
   }
}

// sind jedoch der Titel und die Kategorie undefiniert, wird nur
// nach dem Text gesucht

if($patternText != "#()#i" && $patternTitle == "#()#i" &&
$category == "0") {
   foreach($body as $weitarr) {
       foreach($weitarr as $text) {
           if (@preg_match($patternText, strtolower($text))) {
               if(!array_key_exists($file,$such_array)) {
                  $such_array[$file]=$xmlFile;
                  $counter++;
               }
           }
       }
   }
}
?>
```

Die Ergebnisse der Suche werden als SimpleXML-Objekt mit dem Dateinamen als Schlüssel ins `$such_array` gespeichert. Um die Handhabung später zu vereinfachen wird die Funktion `simplexml2array()` aus der inkludierten Datei „functions.inc.php“ angewandt und zusätzlich das Array noch serialisiert. Das hat den Vorteil, dass durch die PHP-Funktion `serialize()` das Array in eine Zeichenkette umgewandelt wird, die aber die Struktur des Arrays beibehält:

```
<? echo serialize(simplexml2array($such_array)); ?>
```

Die Ausgabe mit `echo` erfolgt, damit das serialisierte Array im „apabasis.php“ mit der PHP-Funktion `file_get_contents()` verwendet werden kann.

6.2.4.5 dom.php

Das Script „dom.php“ wird nach Aufruf aus der Datei „router.php“ aufgerufen und ist physisch am Eingangsserver abgelegt. Es fragt zuerst nach dem übergebenen Ausdruck ab, fehlt dieser, wird das Script abgebrochen. Danach wird geprüft ob die angeforderte Datei existiert. Ist dies der Fall, lädt es nach Übernahme des Dateinamens per GET-Methode die entsprechende Datei mit DOM (Document Object Model). Ansonsten wird ein Mail mit der Fehlermeldung versendet. Per XPath werden die benötigten Tags aus dem NewsML-Dokument ausgelesen. Sollte ein Tag nicht vorhanden sein, wird ein leeres Ergebnis zurückgeliefert. So ist die Ausführung des Scripts nicht gefährdet, da das Script nicht abgebrochen wird und auch keine Fehlermeldung ausgegeben wird, sondern lediglich die zu befüllende Variable leer bleibt.

```
<?
$pfad = "/apa/basis/";
$filename = $_GET["file"];

if(file_exists($pfad.$filename)) {
      $doc=domDocument::load($pfad.$filename);

      $xpath = new DOMXPath($doc);

      // um den XHTML-Bereich im NewsML für XPath verfügbar zu
      // machen, muss er mit registerNamespace angegeben werden

      $xpath->registerNamespace("x","http://www.w3.org/1999/
      xhtml");

      $hl_query = "//HeadLine";
      $headline = $xpath->query($hl_query)->item(0)->nodeValue;

      $shl_query = "//SubHeadLine";
      $subheadline = $xpath->query($shl_query)->item(0)
                     ->nodeValue;

      $lc_query = "//Location/Property/@Value";
      $location = $xpath->query($lc_query)->item(0)->nodeValue;
```

```
$b_query = "//DataContent/x:html/x:body";
$body = $xpath->query($b_query)->item(0)->nodeValue;

$i_query = "//Instruction/@FormalName";
$instruction = $xpath->query($i_query)->item(0)->nodeValue;

$s_query = "//Source/Party";
$source = $xpath->query($s_query);
$z = 0;

// da es möglich ist, dass es mehrere Sourcen (Quellen)
// gibt, wird dieser Bereich mit einer foreach-Schleife
// durchlaufen

foreach ($source as $teile) {
     if($teile->hasAttributes()) {
          foreach ($teile->attributes as $attribute){
               if($z != 0) $sourcen.=", ";
           $sourcen .= $attribute->value;
           $z++;
       }
     }
}

$d_query = "//DateAndTime";
$date = $xpath->query($d_query)->item(0)->nodeValue;

$c_query = "//Creator/Party";
$creator = $xpath->query($c_query);
$z = 0;

// da es möglich ist, dass es mehrere Creator (Verfasser)
// gibt, wird dieser Bereich mit einer foreach-Schleife
// durchlaufen

foreach ($creator as $teile) {
     if($teile->hasAttributes()) {
          foreach ($teile->attributes as $attribute){
               if($z != 0) $creatorn.=", ";
           $creatorn .= $attribute->value;
           $z++;
       }
     }
}

$kwl_query = "//KeywordLine";
$keywordline = $xpath->query($kwl_query)->item(0)
                ->nodeValue;

$sc_query = "//SubjectCode/Subject";
$newscode = $xpath->query($sc_query);
$z = 0;

// da es möglich ist, dass es mehrere Subjects (Kategorien)
```

```
      // gibt, wird dieser Bereich mit einer foreach-Schleife
      // durchlaufen

      foreach ($newscode as $teile) {
            if($teile->hasAttribute("Vocabulary")) {
                  foreach ($teile->attributes as $key =>
                  $attribute){
                        if($key=="FormalName") {
                              if($z != 0) $newscoden.=", ";
                         $newscoden .= $attribute->value;
                         $z++;
                        }
                  }
            }
      }

      $as_query = "//AssociatedWith/@NewsItem";
      $associated = $xpath->query($as_query)->item(0)->nodeValue;

      $li_query = '//Metadata/Property[@FormalName="Homepage"]/
                  @Value';
      $link = $xpath->query($li_query)->item(0)->nodeValue;

      $cl_query = "//CopyrightLine";
      $copyline = $xpath->query($cl_query)->item(0)->nodeValue;

      $l_query = "//LabelText";
      $label = $xpath->query($l_query)->item(0)->nodeValue;

      echo $headline.'<bez>'.$subheadline.'<bez>'.$location.
      '<bez>'.$body.'<bez>'.$sourcen.'<bez>'.$date.'<bez>'.
      $creatorn.'<bez>'.$keywordline.'<bez>'.$newscoden.'<bez>'.
      $associated.'<bez>'.$link.'<bez>'.$copyline.'<bez>'.$label.
      '<bez>'.$instruction;
} else {
      mail("bettina.tuller@kurier.at", "Fehler in dom.php", "das
      angeforderte File ".$filename." konnte nicht gefunden und
      geladen werden");
}
?>
```

Die Ausgabe mit `echo` erfolgt, damit die Zeichenkette im „router.php“ mit der PHP-Funktion `file_get_contents()` verwendet werden kann.

6.2.4.6 parsertopic.php

Die Datei „parsertopic.php“ wird aus dem Script „apabasis.php“ aufgerufen und stellt lediglich die APA-Kategorien bereit in welche die Mitteilungen eingeordnet sind. Dies geschieht über das Laden mit der PHP-Funktion

`simplexml_load_file()`. Schlägt dies fehl, wird auf eine lokale Sicherung der APA-Kategorien zurückgegriffen und ein Mail mit entsprechender Fehlermeldung versendet. Danach wird das geladene SimpleXML-Objekt mit der Funktion `simplexml2array()`, aus dem inkludierten „functions.inc.php“, in ein Array und mit der PHP-Funktion `serialize()` in eine Zeichenkette umgewandelt.

```
<?
include("functions.inc.php");

if($xmltopic = simplexml_load_file("http://www.apa-it.at/NewsML/
IPTC/topicset/topicset.apa-cat.xml")) {

} else {
      $xmltopic = simplexml_load_file("{Pfad}/topicset.apa-
      cat.xml");
      mail("bettina.tuller@kurier.at", "Fehler in
      parsertopic.php", "Fehler beim Laden der APA-Kategorien! Bin
      auf das lokal gespeicherte File ausgewichen, überprüf
      mich!");
}
echo serialize(simplexml2array($xmltopic));
?>
```

6.2.4.7 apabasis.php

Die Datei „apabasis.php“ wird über das neu erstellte Netautor-Layout inkludiert (siehe Abbildung 6). Hier wird mit `header('Content-Type: text/html; charset=utf-8')`der Header definiert, um die richtige Darstellung von Sonderzeichen im „Media Center“ zu gewährleisten. Als nächstes wird die Datei „functions.inc.php“ inkludiert, welche alle benötigten PHP-Funktionen enthält, und der verschlüsselte Ausdruck zur Absicherung generiert.

```
<?
header('Content-Type: text/html; charset=utf-8');
include("functions.inc.php");
$var = md5("{Ausdruck}");
?>
```

Die JavaScript Funktion `popup()` öffnet die referenzierte Datei „router.php“ in einem Popup-Fenster.

```
<script type="text/javascript">
function popup (link) {
  fenster = window.open(link, "Popup", "scrollbars=yes");
  fenster.focus();
}
</script>
```

Die Ausgabe der Datei „parsertopic.php“ wird mit der PHP-Funktion `file_get_contents()` geladen. Da die Ausgabe in der Datei „parsertopic.php“ mit der PHP-Funktion `serialize()` in eine Zeichenkette umgewandelt wurde, muss sie hier nun wieder in ein Array zurück umgewandelt werden, mit der Funktion `unserialize()`. Gleichzeitig wird überprüft ob die Kategorien geladen werden können.

```
<?
if($unser_topic = unserialize(file_get_contents("{Pfad}/
parsertopic.php"))) {
?>
```

Im nächsten Schritt werden die APA-Kategorien aus dem Array ausgelesen und in `$category_arr` gespeichert. Der Standard-Wert „--select--“ wird dann mit der PHP-Funktion `array_unshift()` vor allen bereits eingetragenen Werten als erster Wert eingetragen.

```
<?
      $subject_arr = $unser_topic["NewsItem"]["TopicSet"]
                     ["Topic"];

      foreach($subject_arr as $arr_im_arr) {
            $category_arr[$arr_im_arr["FormalName"]] =
            $arr_im_arr["Description"][0];
      }

      array_unshift($category_arr, "--select--");
      $categories = 1;

} else {
      $categories =0;
      mail("bettina.tuller@kurier.at","Fehler in
      apabasis.php","Fehler beim Übernehmen der APA-Kategorien,
      ich konnte kein File laden, hilf mir!");
}
?>
```

Nun wird das Formular welches die Suchkriterien festlegt erstellt. Können die APA-Kategorien nicht geladen werden, wird das Drop-Down Menü ausgeblendet.

```
<form action="" method="POST" name="formapa">
<font class="text">Titel<br></font>
<input name="titelsuche" type="CHAR" class="input150" value="<?
echo $_REQUEST["titelsuche"]; ?>">
<font class="text">Volltext<br></font>
<input name="textsuche" type="CHAR" class="input150" value="<?
echo $_REQUEST["textsuche"]; ?>">

<? if ($categories) { ?>
     <font class="text">Kategorien<br></font>
     <select name="category" class="input150">
     <?
     foreach($category_arr as $key => $val) {
          echo "<option ";
          if($_REQUEST["category"]==$key) echo "selected ";
          echo "value='".$key."'>".$val."</option>";
     }
     ?>
     </select>
<? } ?>

<input type="submit" name="npf_search" value="Suchen"
class="search" />
```

Der nachfolgende Bereich soll nur dann sichtbar sein, wenn eine Suchanfrage durchgeführt wurde.

```
<? if ($_POST["npf_search"] == "Suchen") { ?>
```

Da in den Textfeldern für die Titel- und Textsuche nicht nur ein Wort stehen kann, sondern auch mehrere, wird die Zeichenkette bei jedem Leerzeichen, mit der PHP-Funktion `explode()`, getrennt. Stattdessen wird ein spezieller Separator eingefügt, der so jedes Suchwort vom anderen trennt.

```
<?
$such_title = explode(" ",strip_tags(trim(
               $_REQUEST["titelsuche"])));
$anzahl_title = count($such_title) - 1;

foreach ($such_title as $tkey => $tval) {
     $patternTitle .= strip_tags(strtolower($tval));
     if ($tkey < $anzahl_title) $patternTitle .= "|";
}
```

```
$such_text = explode(" ",strip_tags(trim(
              $_REQUEST["textsuche"])));
$anzahl_text = count($such_text) - 1;

foreach ($such_text as $tkey => $tval) {
     $patternText .= strip_tags(strtolower($tval));
     if ($tkey < $anzahl_text) $patternText .= "|";
}
?>
```

Die Datei „apabasis_simplexml.php“ wird mit den generierten Suchbegriffen und dem verschlüsselten Ausdruck aufgerufen. Die Ausgabe wird mit der PHP-Funktion `file_get_contents()` geholt und mit der PHP-Funktion `unserialize()` in ein Array zurück umgewandelt. Kann der Inhalt nicht geladen werden wird ein Mail versendet.

```
<?
if($such_array = unserialize(file_get_contents("{Pfad}/
apabasis_simplexml.php?patterntext=".$patternText."
&patterntitle=".$patternTitle."&category=".$_REQUEST["category"].
"&meinvar=".$var))) {

                {nachfolgender Inhalt}

} else {
     echo "Fehler!!!<br><br>Bleiben Sie ruhig, Hilfe ist
     unterwegs!<br>Es wurde ein Mail an den Administrator
     versendet!";

     mail("bettina.tuller@kurier.at", "Fehler in apabasis.php",
     "Fehler beim Laden des such_array aus
     apabasis_simplexml.php. Überprüf mich!");
}
?>
```

Nun werden für jeden Eintrag im `such_array` die benötigten Informationen Datum (`filedatum`), Dringlichkeit (`urgency`), Titel (`title`), Kurztext (`kurztext`), Inhalt (`content`), Sourcen (`credit`) und Ort (`location`), ermittelt.

```
<?
foreach ($such_array as $skey => $sucherg) {
     $filedatum = $sucherg["NewsEnvelope"]["DateAndTime"];
     $urgency = $sucherg["NewsItem"]["NewsManagement"]
                ["Urgency"]["@attributes"]["FormalName"];
```

```
     $title = $sucherg["NewsItem"]["NewsComponent"]
               ["ContentItem"]["DataContent"]["html"]["head"]
               ["title"];
     $kurztext = $sucherg["NewsItem"]["NewsComponent"]
               ["NewsLines"]["SubHeadLine"];
     $content = $sucherg["NewsItem"]["NewsComponent"]
               ["ContentItem"]["DataContent"]["html"]["body"];
     $credit = $sucherg["NewsItem"]["NewsComponent"]
               ["AdministrativeMetadata"]["Source"];
     $location = $sucherg["NewsItem"]["NewsComponent"]
               ["DescriptiveMetadata"]["Location"]["Property"]
               ["@attributes"]["Value"];
?>
```

Durch Klicken des Titels öffnet sich das Popup-Fenster zur Voransicht. Hier werden an die Datei „router.php“ der Dateiname, der verschlüsselte Ausdruck und die Suchanfragen übermittelt.

```
<?
echo "<a href='router.php?file=".$skey."&meinvar="
     .$var."&?patterntext=".$patternText."&patterntitle="
     .$patternTitle."&category=".$_REQUEST["category"]."'
     style='color:".textfarbe($urgency)."' class='textboldbig'
     onclick='popup(this.href); return false'>";

// wird ein Suchwort im Titel gefunden, wird es an die Funktion
// „highlight“ geschickt, welche es farblich hinterlegt

if($patternTitle!="") {
          echo highlight($patternTitle,$title);
} else {
     echo $title;
}
echo "</a>";

// Datum und Uhrzeit der Meldung anzeigen

echo "<font class='text'>".substr($filedatum,6,2)."."
.substr($filedatum,4,2).".".substr($filedatum,0,4)." - "
.substr($filedatum,9,2).":".substr($filedatum,11,2).":"
.substr($filedatum,13,2)."<br></font>
?>
```

Da es durchaus mehrere Sourcen, sowie mehrere Absätze im Inhaltsbereich, geben kann, werden diese beiden Bereiche mit einer `foreach`-Schleife durchlaufen.

```
<?
if(is_array($credit["Party"])) {
   foreach($credit["Party"] as $key => $value) {
       if($anzcredit==1) {
           $credits .= $value["FormalName"];
       } else {
            if($key==0) {
               $credits .= $value["@attributes"]["FormalName"];
            } else {
                $credits .= ", ".$value["@attributes"]
               ["FormalName"];
            }
       }
   }
} else {
   $credits = $credit;
}

$contents = $location;

foreach($content as $schl => $weitarr) {
    if(is_array($weitarr)) {
       foreach($weitarr as $text) {
           $contents.= "<".$schl.">".$text."</".$schl.">\r\n";
       }
    } else {
        $contents .= "<".$schl.">".$weitarr."</".$schl.">\r\n";
    }
}
?>
```

Durch Klicken des Buttons „Erstellen“ werden per JavaScript die Inhalte in die CMS-Maske übernommen.

```
<?
<a href=\"#\" onclick=\"parent.edit.document.form1.
elements['xmlval_TITLE[0]'].value ='".repl($title)."';
parent.edit.document.form1.elements['xmlval_SUMMARY[0]'].value =
'". repl($kurztext)."';
parent.edit.document.form1.elements['xmlval_CONTENT[0][C_TEXT]
[0]'].value = '". repl($contents)."';
parent.edit.document.form1.elements['xmlval_CREDIT[0]'].value =
'". repl($credits)."';\">
<img src='img/erstellen.gif' border='0' title='Artikel
erstellen'></a>
?>
```

6.2.4.8 router.php

Die Datei „router.php“ stellt die Voransicht des gewählten Artikels dar. Sie wird in der Datei „apabasis.php“ über einen Klick auf den Titel aufgerufen und ist am Verwaltungsserver abgelegt. Um die Darstellung von Sonderzeichen zu gewährleisten wird zuerst der `header` deklariert. Weiters wird wieder nach dem verschlüsselten Ausdruck abgefragt und die Datei „functions.inc.php“ inkludiert. Weiters wird die per GET-Methode übergebene Variable in `akt_file` gespeichert.

```
<?
header('Content-Type: text/html; charset=utf-8');
if($_GET["meinvar"] != md5("{Ausdruck}")) {
      echo "nicht berechtigt";
      exit();
}
include("functions.inc.php");
$akt_file = $_GET["file"];
?>
```

Die darzustellenden Inhalte werden über die Datei „dom.php“ mit dem Namen der aktuellen Datei und dem verschlüsselten Ausdruck aufgerufen. Hier werden die Inhalte wiederum mit der PHP-Funktion `file_get_contents()` geholt. Ist dies nicht möglich, wird ein Mail mit einer Fehlermeldung versendet. Weiters wird die erstellte Zeichenkette bei einem bestimmten Ausdruck mit der PHP-Funktion `explode()` geteilt.

```
<?
if($mlneu = file_get_contents("{Pfad}/dom.php?file="
.$akt_file."&meinvar=".$var)) {
      $boom = explode("<bez>",$mlneu);
?>
```

Als nächstes wird jeder erhaltene Wert der Funktion `repl()` übergeben, um bestimmte Zeichen, die möglicherweise Fehler verursachen können oder die Darstellung beeinträchtigen, zu ersetzen. Weiters werden die erhaltenen Werte aus dem Array an Variablen mit aussagekräftigem Namen zugewiesen.

```
<?
foreach ($boom as $key => $teil) {
     $boom[$key] = repl($teil);
}

$title = $boom[0];
$kurztext = $boom[1];
$location = $boom[2];
$content = $boom[3];
$credit = $boom[4];
$date = $boom[5];
$creator = $boom[6];
$keyword = $boom[7];
$newscode = $boom[8];
$asso = $boom[9];
$webtipp = $boom[10];
$copyline = $boom[11];
$label = $boom[12];
$instruction = $boom[13];

if($location != "") {
     $content = $location." - ".$content;
}
?>
```

Nun werden die gewonnenen Inhalte dargestellt.

```
<?
echo "<html><head><title>".$title."</title>
<link rel='stylesheet' type='text/css' href='style.css'>
</head><body><h2>".$title."</h2><h4>".$kurztext."</h4>";

// wenn es Anweisungen gibt, wie z.B. fortsetzung-moeglich,
// werden sie nach dem Inhaltsbereich angezeigt

if($instruction != "") {
     $content .= "(".$instruction.")";
}

echo "<p>".$content."</p>
<p>Datum: ".substr($date,6,2).".".substr($date,4,2)."."
.substr($date,0,4)." - ".substr($date,9,2).":"
.substr($date,11,2).":".substr($date,13,2)."<br>
Creator: ".$creator."<br>\r\nSource: ".$credit."<br><br>
\r\nKeywords: ".$keyword."<br>\r\nAPARessort: ".$newscode;

// wenn ein assoziiertes File in dem NewsML-Dokument eingetragen
// ist, wird der Dateiname und das Datum der Meldung eruiert

if ($asso != "") {
     $asso = explode(":",$asso);
     $assoc = substr($asso[4],8,7)."-".substr($asso[4],0,8);
```

```
      $edate = substr($asso[4],0,8);
      $jahr = substr($edate,0,4);
      $monat = substr($edate,4,2);
      $tag = substr($edate,6,2);

      $tmstmp = mktime(0,0,0,$monat,$tag,$jahr);
      $max = time()-(60*60*24*1);

      // damit nicht eine ältere Datei geladen wird, welche
      // vielleicht gar nicht mehr existiert, wird der timestamp
      // der Datei überprüft und mit der Zeiteinstellung der
      // Löschroutine für APA-Dateien verglichen -> $max

      if($tmstmp>=$max) {
            $associatedfile = $assoc.".xml";
            $headlineasso = file_get_contents("{Pfad}/dom.php?
            file=".$associatedfile."&meinvar=".$var);
            $booma = explode("<bez>",$headlineasso);
      }
      echo "<br><br>\r\nVerwandter Artikel: <a href='{Pfad}/
      router.php?file=".$associatedfile."&meinvar=".$var."'
      target='_blank'>".$booma[0]."</a>";
}

if($webtipp != "") {
      echo "<br><br>\r\nWebtipp: <a href='".$webtipp."'
      target='_blank'>".$webtipp."</a>";
}

echo "<br><br>\r\nCopyright: ".$copyline."<br>\r\nAPA-Nr.:
".$label."</p></body>\r\n</html>";
?>
```

Am Ende im else-Zweig der Anweisung die Initiierung der Fehlermail.

```
<?
} else {
      echo "Fehler!!!<br><br>Bleiben Sie ruhig, Hilfe ist
      unterwegs!<br>Es wurde ein Mail an den Administrator
      versendet!";

      mail("bettina.tuller@kurier.at", "Fehler in router.php",
      "Fehler beim Laden der Inhalte aus dom.php. Überprüf
      mich!");
}
?>
```

6.2.5 Diagramme

Ein Komponentendiagramm zeigt alle Komponenten eines Systems auf. Für das KURIER Online CMS mit Einbindung der NewsML-Dokumente sieht dieses so aus:

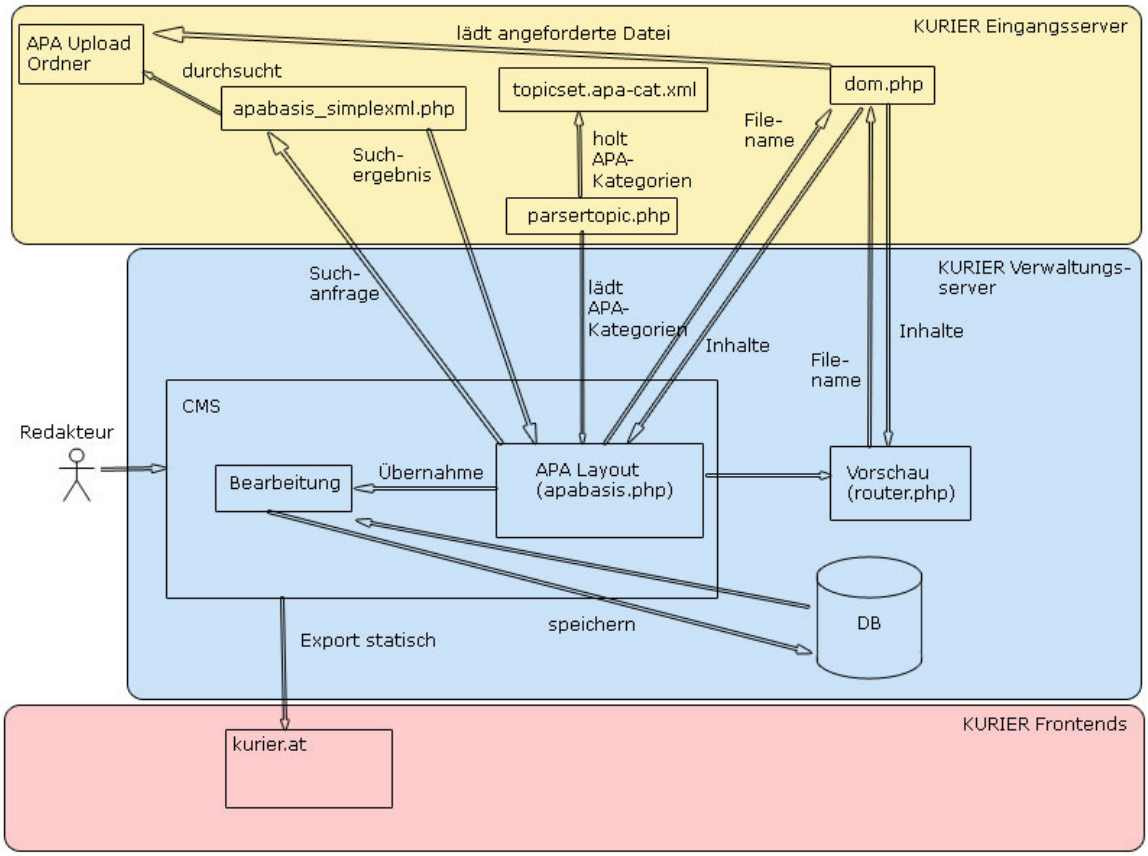

Abbildung 11: Komponentendiagramm vom KURIER Online CMS

Ein weiteres Diagramm namens Use-Case-Diagramm soll die Anwendung eines Produktes vereinfacht darstellen, deswegen trägt es ebenso die Bezeichnung „Anwendungsfalldiagramm“. Nachfolgend ist ein solches Diagramm zum Aufbau und zur Benutzung des „NewsML by APA“ im KURIER Online CMS dargestellt. Hierzu muss angemerkt werden, dass sich im Anwendungsfall nicht viel geändert hat, bis auf die Möglichkeit der Voransicht. Was jedoch anders ist, ist die Herkunft des Artikels: Üblicherweise wird ein Artikel aus der Datenbank geladen, nun kommen die Inhalte direkt aus dem NewsML-Dokument.

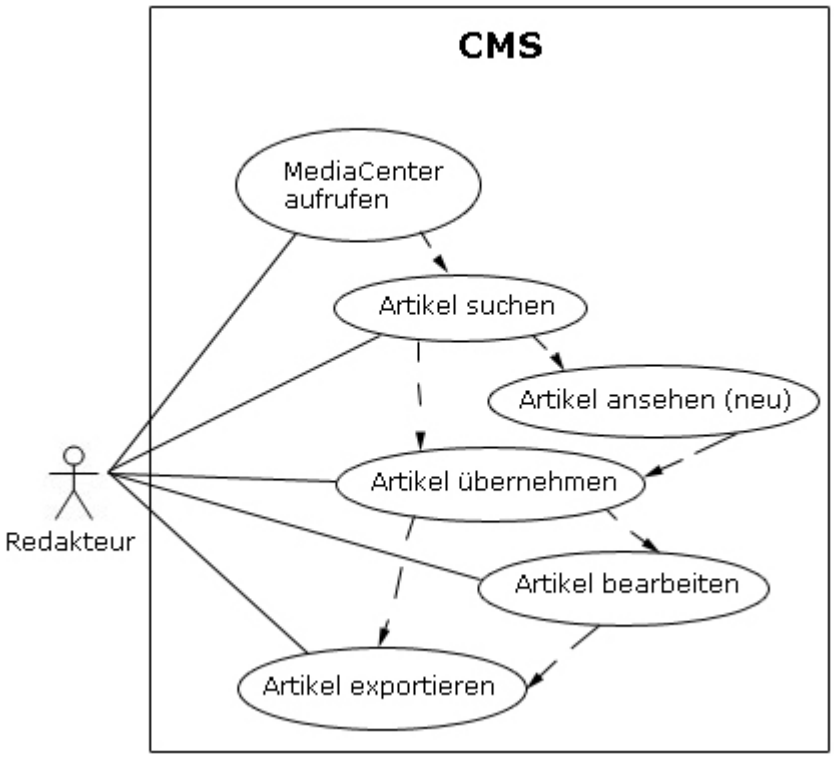

Abbildung 12: Use-Case-Diagramm zur Benutzung von NewsML im CMS

7. Ausblick

Seitens der IPTC wird zurzeit schon fleißig am Nachfolger NewsML 2 Architecture (NAR) gebastelt. Laut den Informationen die auf der Website preisgegeben werden wurde im Mai dieses Jahres bereits die Testphase von NewsML 2 eingeleitet. Die Veröffentlichung soll noch Ende 2006 stattfinden.

Diese Weiterentwicklung von NewsML soll noch mehr Metadaten zur Verfügung stellen und weiterhin auf dem Standard von XML beruhen. Wie sein Vorgänger soll NewsML 2 in der Lage sein mehrere NewsItems und mehrere Medientypen in einem Dokument zu transportieren. Die neue Version soll allerdings die bereits in NewsML 1.x verwendeten „NewsCodes“ der IPTC stärker nutzen und auch in der Lage sein mit anderen IPTC-Standards, wie z.B. SportsML, besser interagieren. Es wird bessere Möglichkeiten zum Managen der Nachrichten geben, so können z.B. die Vertreiber wählen ob sie den ganzen Umfang an zur Verfügung gestellten NewsML-Tags verwenden oder gar nur eine ausgewählte Teilmenge davon.[33] Neu ist, dass es PackageItems geben soll, in denen NewsItems, TopicItems und auch andere PackageItems enthalten sein können.[34] Das NewsItem enthält dabei die eigentliche, aktuelle Nachricht, während im TopicItem Hintergrundwissen zu diesem Thema enthalten ist.

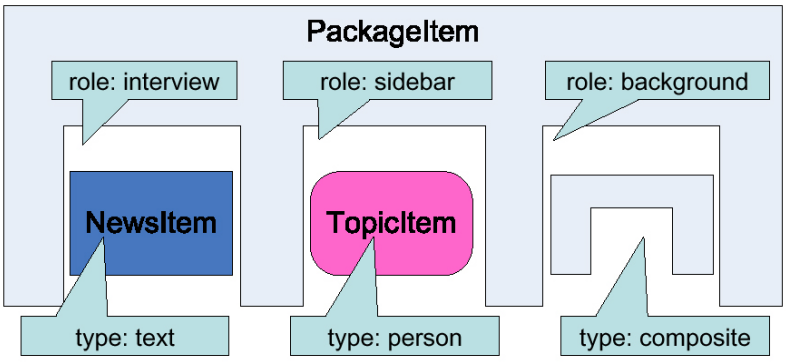

Abbildung 13: PackageItem aus NAR

[33] vgl. „What’s NewsML 2?“ von der IPTC, online unter <http://www.iptc.org/pages/index.php> (26.06.2006)

[34] vgl. „IPTC News Architecture“ von der IPTC, online unter <http://iptc.org/pdl.php?fn=20050708-IPTC-NewsArch4W3C.pdf> (26.06.2006)

Hier ein Auszug eines Beispiels der Struktur von einem TopicItem in NewsML 2:

```
<?xml version="1.0" encoding="UTF-8"?>
<topicItem xmlns="http://iptc.org/std/newsml/2006-05-01/">
   <catalog>
      <scheme alias="ccls"
      uri="http://iptc.org/newscodes/contentclass#"/>
      <scheme uri="http://iptc.org/newscodes/namerole#"
      alias="nrol"/>
      <scheme uri="http://en.wikipedia.org/wiki/" alias="enwiki"/>
   </catalog>
   <itemMeta>
      <contentClass code="ccls:person"/>
      <provider literal="IPTC"/>
      <itemCreated>2006-05-24T20:00:00Z</itemCreated>
      <released>2006-05-24T20:01:00Z</released>
      <title>Wolfgang Amadeus Mozart</title>
      <copyrightNotice>Copyright 2006 IPTC</copyrightNotice>
   </itemMeta>
   <contentMeta>
      <contentCreated>2006-05-24</contentCreated>
      <creator literal="Michael Steidl"/>
      <infoSource code="enwiki:Mozart"/>
   </contentMeta>
   <personDefinition>
      <name role="nrol:wellknown">Wolfgang Amadeus Mozart</name>
      <name role="nrol:baptised">Joannes Chrysostomus Wolfgangus
      Theophilus Mozart</name>
      <definition>Mozart is among the most significant and
      enduringly popular composers of European classical music.
      His enormous output includes works that are widely
      acknowledged as pinnacles of symphonic, chamber, piano,
      operatic, and choral music.</definition>
      <born>1756-01-27</born>
      <died>1791-12-05</died>
      <occupation literal="music composer"/>
      <skill literal="pianist"/>
   </personDefinition>
</topicItem>
```

Eine andere Neuerung soll es auch bei KURIER Online geben. Bis jetzt bezieht KURIER Online das NewsML-Format ausschließlich von der APA, welche ihren Basisdienst (Text) und Grafikdienst zur Verfügung stellt. Mittlerweile werden für kurier.at Streamingvideos per Flash 8 angeboten, die von der Nachrichtenagentur Reuters bezogen werden. Allerdings stehen derzeit noch keine Metadaten zu den Videos zur Verfügung. Laut Technikabteilung des KURIER Online soll sich dies allerdings noch im Laufe dieses Jahres ändern und die Belieferung auf Videos im NewsML-Format umgestellt werden.

8. Zusammenfassung

Die Erweiterung des KURIER Online CMS um die Schnittstelle „NewsML zu CMS“ gestaltete sich etwas schwieriger als zunächst angenommen. Durch die Rücksichtnahme auf die gegebene Serverumgebung war die Umsetzung etwas eingeschränkt, aber nicht behindert. So wurden andere Mittel und Wege gefunden, welche dann zum gewünschten Ergebnis führten.

Die gelieferten Nachrichtenartikel im NewsML-Format werden durch die Verwendung von sechs erstellten Scripts automatisch übernommen und stehen so direkt im KURIER CMS zur Verfügung. Die PHP-Scripts sind, je nach Funktionalität, auf die beiden interagierenden Server aufgeteilt. Am Verwaltungsserver wird die Datei „apabasis.php“ direkt im CMS aufgerufen. Hier werden die Suchoptionen angegeben, unter anderem die Kategorie, die aus der Datei „parsertopic.php“ geholt wird. Die eingegebenen Suchkriterien werden dann der Datei „apabasis_simplexml.php“ am Eingangsserver übergeben. Dieses Script durchsucht alle NewsML-Dokumente die den Status „Usable“ aufweisen nach den übergebenen Suchkriterien. Trifft die Suchanfrage zu, werden der Name und der Inhalt der Datei in ein Array gespeichert, welches dann in „apabasis.php“ ausgelesen und angezeigt wird.

Klickt man auf den Titel hat man über ein Popup-Fenster die Möglichkeit zur Voransicht, welche mit der Datei „router.php“ angezeigt wird. Hier wird die Datei „dom.php“ aufgerufen, welcher bei Aufruf der Dateiname übergeben wird und die benötigten Tags ausgelesen werden. Diese werden in einer Zeichenkette gespeichert, welche dann wiederum in „router.php“ ausgelesen und zur Anzeige verarbeitet wird. Klickt man bei den Suchergebnissen jedoch nicht auf den Titel, sondern auf den Button „Erstellen“, wird der Artikel in die Maske des CMS übernommen und kann wie ein normaler Artikel bearbeitet, gespeichert und exportiert werden, um dann online zu erscheinen.

Die Performance des Prozesses verbesserte sich mit der Weiterentwicklung der Ideen. Während die erste Lösung mit der Übergabe des gesamten Arrays den Höchstwert von 12 und Minimalwert von 3 Sekunden zur Anzeige benötigte, bleibt sie nun immerhin, je nach Suchanfrage, zwischen einem Minimalwert von 100 Nanosekunden und einem Maximalwert von 1,5 Sekunden.

Das „Error Handling“ umfasst die Überprüfung der empfangenen Dateien. So werden die ausgelesenen APA-Kategorien überprüft und bei Auftreten eines Fehler die lokale Sicherung verwendet. In jedem Fehlerfall wird ein Mail mit entsprechender Meldung an den Administrator versendet. Weiters wird auch die Erstellung des Arrays in „apabasis_simplexml.php“ überprüft. So wird jede Datei unter anderem auf Dateigröße überprüft bevor sie übernommen wird, um sicherzugehen, dass keine fehlerhaften Dateien mit der Größe von 0 Byte übernommen werden. Außerdem wird das Array mit den gespeicherten Dateinamen geprüft bevor es mit einer Schleife durchlaufen wird um die Inhalte zu parsen, wobei beim Laden der Datei wiederum eine Überprüfung stattfindet, ob dies fehlerfrei geschieht. Im Script „dom.php“ wird vor dem Laden nach der angeforderten Datei abgefragt. In der Datei „router.php“ wird geprüft ob die Inhalte aus „dom.php“ übernommen wurden. In „apabasis.php“ werden die Kategorien und die Übernahme des Arrays mit den Suchergebnissen überprüft.

Es bleibt abzuwarten ob sich die Integration von „NewsML by APA“ gegenüber dem APA-OnlineManager durchsetzen kann. Fakt ist jedoch, dass es durchaus eine Arbeitserleichterung darstellt. Im Arbeitsablauf eines Online-Redakteurs ist das „Copy / Paste“ – Verfahren alltäglich, aber nervenaufreibend und auch zeitraubend. Hier soll die direkte Einbettung und Übernahme der Meldungen Erleichterung schaffen. Abgesehen vom Design an der Benutzeroberfläche unterscheidet sich die Darstellung einer Meldung im AOM nicht wesentlich von der Implementierung im KURIER-CMS, welche jedoch den Vorteil hat, dass die Artikel direkt übernommen und bearbeitet werden können. So sind die Redakteure nicht an ihren Arbeitsplatz gebunden und können, egal wo sie sich gerade befinden, wichtige Nachrichten am schnellsten Weg verarbeiten.

9. Quellenverzeichnis

1. Eric van der Vlist, XML-Schema, O'Reilly 2003

Internetrecherche:

2. http://www.w3c.de
3. http://www.newsml.org
4. http://about.reuters.com
5. http://www.iptc.org
6. http://de.wikipedia.org/wiki
7. http://www.apa-it.at
8. http://www.nitf.org
9. http://www.prismstandard.org
10. http://www.ipsi.fraunhofer.de/merit/lehre/veranstaltungen/intercosmical2006/material/InterCOSMICAL2006_3.I.1_NewsML_OezguerDagdelen.pdf
11. http://www.rss-verzeichnis.de/alles-ueber-rss.php
12. http://relaxng.org/
13. http://www.gungfu.de/facts/archives/2005/07/29/mein-xmlhttprequest-und-ich/

Abbildungen:

1. http://www.apa-it.at/NewsML/Deutsch/NewsMLbyAPA-TechDokumentation_v1.06.pdf
2. http://about.reuters.com/newsml/newsml.ppt
3. http://www.apa-it.at/NewsML/Deutsch/NewsMLbyAPA-TechDokumentation_v1.06.pdf

13. http://iptc.org/pdl.php?fn=20050708-IPTC-NewsArch4W3C.pdf

10. Glossar

AJAX: „Asynchronous JavaScript and XML“ arbeitet nach dem Prinzip der asynchronen Datenübertragung, was unter anderem eine HTTP-Anfrage innerhalb einer Website ermöglicht, ohne die Website neu laden zu müssen (siehe auch XMLHttpRequest).
(vgl. http://developer.mozilla.org/de/docs/AJAX:Getting_Started)

APA-OnlineManager: Der APA-OnlineManager (AOM) stellt die Informationsquelle der KURIER Online-Redakteure dar. Diese Anwendung ermöglicht das rasche Abrufen von Informationen samt Bildern, Grafiken und Audiodateien, welche von der APA bereitgestellt werden.
(vgl. http://www.apa-it.at/)

Array: engl.: Anordnung, Aufstellung, Reihe, Reihung. Ein Array ist eine geordnete Abbildung, welches Werte und dazugehörige Schlüssel enthält. Über den Schlüssel ist jeder Wert im Array ansprechbar. Er kann vom Typ „integer“ (eine Zahl) oder „string“ (eine Zeichenkette) sein. Werden keine Schlüssel definiert, werden alle Einträge (Werte) des Arrays mit Zahlen -Schlüsseln, startend bei 0, versehen. Es ist auch möglich ein Array als Wert zu definieren, also mehrere Arrays in einander zu verschachteln. So können Baumstrukturen dargestellt werden. Man spricht dann von einem „multidimesionalen“ Array.
(vgl. http://php.net/)

Atom: Eine Weiterentwicklung von RSS, welche ebenfalls auf XML basiert. Wird vor allem im Bereich von Weblogs verwendet, da es plattformunabhängig und einfach zu erweitern ist.

CMS: Ein Content-Management-System (CMS) ist eine Anwendung, welches es ermöglicht den Inhalt einer Website getrennt vom Design und dem Programmierungscode zu bearbeiten.

Dokumenttypdefinition: engl. Document Type Definition (DTD), ist eine Deklaration, die die Struktur eines XML-Dokuments festlegt, indem es die Anwendung von Auszeichnungen beschreibt.

DOM: Das „Document Object Model“ ist eine Schnittstelle die es z.B. ermöglicht dynamisch den Inhalt, aber auch die Struktur einer Website zu aktualisieren. Es wurde vom W3C definiert und arbeitet objektorientiert.

Enkodierung: Unter einer Enkodierung versteht man die Kodierung von Zeichen. Der angegebene Zeichensatz sorgt für die richtige Darstellung der Zeichen. In der deutschen Sprache ist vor allem die Darstellung der Umlaute immer wieder problematisch. Diese Missinterpretation kann durch die Angabe des richtigen Kodierungsstandards umgangen werden. Die bekanntesten Kodierungsstandards in unseren Breiten stellen dabei UTF-8, ISO-8859-1 und ISO-8859-15 dar.

Feed: Der Begriff „Feed“ wird aus dem englischen mit „Einspeisung“ oder „Zufuhr“ übersetzt. Feed oder Newsfeed wird im deutschen als Bezeichnung für den Transport von Nachrichten verwendet.

File Transfer Protocol: Das File Transfer Protocol (FTP) stellt ein Protokoll zur Übertragung von Dateien dar. Mit Hilfe dieses Netzwerkprotokolls ist es möglich Dateien über TCP/IP zwischen Server und Client (Download, Upload) aber auch zwischen zwei Servern zu übertragen.

GET: Eine HTTP-Request-Methode um Inhalte vom Server anzufordern. Es gibt noch einige weitere Methoden, die gebräuchlichsten stellen jedoch die Methoden GET und POST dar. Mit der POST-Methode werden nicht nur Daten angefragt, sondern auch an den Server übermittelt.

HTML: Die „Hypertext Markup Language“ stellt eine Auszeichnungssprache dar, um Webinhalte zu strukturieren. XHTML (Extensible Hypertext Markup Language) stellt eine Weiterentwicklung davon dar, die jedoch auf XML basiert.

HTTP: Das „Hypertext Transfer Protocol“ stellt ein Netzwerkprotokoll dar, welches Daten über ein Netzwerk überträgt. Hiermit werden vor allem Websites in den Browser geladen.

IFrame: Inlineframes (iframes) werden in der Regel dafür genutzt um andere Dokumente in eine Website einzubinden. Iframes werden nur allzu häufig mit normalen Frames verwechselt, welche Websites in verschiedene Bereiche unterteilen.

IPTC: Die IPTC (International Press Telecommunications Council) ist ein Konsortium, welches sich aus den weltweit größten Nachrichtenagenturen und Nachrichtenanbietern zusammensetzt. Es entwickelt und wartet technische Standards für einen besseren Nachrichtenaustausch. Der Sitz des Unternehmens ist in Windsor, England.

Komponentendiagramm: Ein Komponentendiagramm dient zur Veranschaulichung einer Struktur eines erstellten Systems. Hier werden alle Komponenten des Systems mit allen Schnittstellen dargestellt.

Namensraum: engl. „namespace“. Dieser stellt einen eigens deklarierten Bereich dar, welcher unter anderem Namenskonflikte verhindert, da jedem Namen noch zusätzlich das Namensraumkürzel vorangestellt wird. So können Elemente und Attribute ein Namensraum-Präfix aufweisen, müssen es aber nicht, da die Verwendung von Namensräumen optional ist.

PHP: PHP steht für „Hypertext Preprocessor“ und ist eine Open-Source-Software, die über zahlreiche Funktionsbibliotheken verfügt. PHP ermöglicht die Erstellung dynamischer Websites oder ganzer Webanwendungen.

ProgramGuideML: ProgramGuideML ist wie NewsML ein spezielles Format, das jedoch in diesem Fall zum Austauschen von Radio- bzw. TV-Informationen benutzt wird. (vgl. http://www.iptc.org/)

RDF: Das „Resource Description Framework“ stellt Metadaten im World Wide Web bereit. Hiermit können nicht direkt enthaltene Informationen einer Website zugänglich gemacht werden, wie etwa Titel oder Copyright.

SportsML: SportsML ist, ähnlich zu ProgramGuideML, ein speziell entwickeltes Format. Es dient unter anderem dazu Spielpläne, Punktestände und Statistiken im Bereich der Sportnachrichten zu übertragen. (vgl. http://www.iptc.org/)

Tag: Es gibt zwei Arten von Tags: Einerseits den normalen Tag, der einen auszuzeichnenden Inhaltsbereich öffnet und der nach Angabe der Informationen wieder geschlossen wird und andererseits den so genannten „Standalone Tag“, welcher allein steht und nicht geschlossen werden muss.

Use-Case-Diagramm: Ein Use-Case-Diagramm veranschaulicht den Anwendungsfall eines Systems. Hier werden die agierenden Personen sowie die Anwendungsfälle samt Beziehungen und Abhängigkeiten dargestellt.

W3C: Das „World Wide Web Consortium“ entwickelt diverse Webtechnologien, Spezifikationen, Guidelines und Tools um das Potenzial des Webs und seinen Technologien voll auszuschöpfen.

XML: Die Extensible Markup Language ist wie HTML eine Auszeichnungssprache und dient zur Erstellung von Dokumenten in Form einer Baumstruktur.

XMLHttpRequest: eine Klasse die die ActiveX-Objekte des Internet Explorers unterstützt und über JavaScript eine Anfrage per HTTP an den Server schickt und verarbeitet (siehe auch AJAX).
(vgl. http://developer.mozilla.org/de/docs/AJAX:Getting_Started)

XML-Schema: XSD (XML-Schema Definition) stellt neben der DTD und Relax NG eine Definition für XML-Dokumente dar. Im Gegensatz zur DTD wird dieses Schema in Form eines XML-Dokuments beschrieben.

XSLT: Mit Hilfe der „Extensible Stylesheet Language Transformations“ ist es möglich XML-Dokumente in eine andere, im XSLT-Dokument festgelegte Struktur umzuformen. Wie der Name aussagt ist es ein „Stylesheet“ für XML, so wie CSS für HTML.

XPath: Über die „XML Path Language“ ist es möglich bestimmte Bereiche eines XML-Dokuments direkt anzusprechen. XPath interpretiert das XML-Dokument als Baum mit dessen Elementknoten, Attributknoten und Textknoten. Auf diese Weise ist jede Auszeichnung samt Inhalt ansprechbar.

11. Abbildungsverzeichnis

Abbildung 1: Übersicht zur NewsML-Struktur ..... - 23 -
Abbildung 2: Beispielaufbau eines NewsItems ..... - 24 -
Abbildung 3: Darstellung von Vererbung der Metadaten ..... - 36 -
Abbildung 4: APA-OnlineManager (Screenshot) ..... - 37 -
Abbildung 5: Das KURIER Online CMS „Digital Workroom“ (Screenshot) ..... - 39 -
Abbildung 6: Hinzufügen eines neuen Layouts (Screenshot) ..... - 40 -
Abbildung 7: Grafische Darstellung der Einbindung des Scripts ins CMS ..... - 41 -
Abbildung 8: Artikelsuche nach Stichwörtern und Kategorie (Screenshot) ..... - 43 -
Abbildung 9: Ansicht eines APA-Artikels im Detail (Screenshot) ..... - 44 -
Abbildung 10: Skizze der Serverlandschaft ..... - 45 -
Abbildung 11: Komponentendiagramm vom KURIER Online CMS ..... - 70 -
Abbildung 12: Use-Case-Diagramm zur Benutzung von NewsML im CMS .... - 71 -
Abbildung 13: PackageItem aus NAR ..... - 72 -

Anhang A: Quelltexte

A.1 functions.inc.php

```
<?
function simplexml2array($xml) {
      if (get_class($xml) == 'SimpleXMLElement') {
            $attributes = $xml->attributes();
            foreach($attributes as $k=>$v) {
                  if ($v) $a[$k] = (string) $v;
            }
            $x = $xml;
            $xml = get_object_vars($xml);
      }
      if (is_array($xml)) {
            if (count($xml) == 0) return (string) $x;
            foreach($xml as $key=>$value) {
                  $r[$key] = simplexml2array($value);
            }
            return $r;
      }

      return (string) $xml;
}

function highlight($pattern,$inhalt) {
      $suche = preg_replace("#(".$pattern.")#i","<span
                  style=\"background-color:#ffbaa2\">
                  \${1}</span>",$inhalt,-1);
      return $suche;
}

function textfarbe($urgency) {
      switch ($urgency) {
            case 1: $color="#CC0000"; //rot
                        break;
            case 2: $color="#EE9900"; //orange
                        break;
            case 3: $color="#0000DD"; //blau
                        break;
            case 4: $color="#00AA00"; //grün
                        break;
            default: $color="#333333"; //dunkelgrau
      }
      return $color;
}
```

```
function repl($var) {
     $var = strip_tags(strtr($var, array("<br>"=>" ",
     "'"=>"&#39;", "\""=>"&#34;", "Utl.: "=>"", "Utl.:"=>"",
     "\n"=>" ")));
     return $var;
}
?>
```

A.2 apabasis_simplexml.php

```
<?
if(!$_GET["meinvar"] == md5("{Ausdruck}")) {
   echo "nicht berechtigt";
   exit();
}

include("functions.inc.php");
$fehler=0;

$patternText= "#(".$_GET["patterntext"].")#i";
$patternTitle= "#(".$_GET["patterntitle"].")#i";
$category= $_GET["category"];

$pfad = "/apa/basis/";
$verz = opendir($pfad);

while (false !== ($file = readdir($verz))) {
   if($file != "." && $file != ".." && ereg('.xml', $file) &&
   filesize($pfad.$file)!=0) {
      $file_arr[filemtime($pfad.$file)] = $file;
   }
}

$anz_files = count($file_arr);

if($anz_files==0) {
     mail("bettina.tuller@kurier.at", "Fehler in
     apabasis_simplexml.php", "keine Files im Array! Bitte
     überprüf mich!");

} else {

krsort($file_arr);
$such_array = array();
$counter=0;

foreach($file_arr as $file) {
   if($xmlFile = simplexml_load_file($pfad.$file)) {
   $status = $xmlFile->NewsItem->NewsManagement->Status[0]
               ->attributes()->FormalName;
```

```
if($status=="Usable") {
   if($counter == 100) break;

   $title = $xmlFile->NewsItem->NewsComponent->ContentItem
            ->DataContent->html->head->title;
   $body = $xmlFile->NewsItem->NewsComponent->ContentItem
           ->DataContent->html->body;
   $subject = $xmlFile->NewsItem->NewsComponent
              ->DescriptiveMetadata->SubjectCode;

   if($category != "0") {
      if($patternTitle != "#()#i" && $patternText != "#()#i") {
         if (($subject->Subject[0]->attributes()->FormalName ==
         $category) && (@preg_match($patternTitle,
         strtolower($title)))) {
              foreach($body as $weitarr) {
                  foreach($weitarr as $text) {
                      if (@preg_match($patternText,
                      strtolower($text))) {
                          if(!array_key_exists($file,$such_
                          array)) {
                              $such_array[$file]=$xmlFile;
                              $counter++;
                          }
                      }
                  }
              }
         }
      } elseif($patternTitle != "#()#i") {
            if (($subject->Subject[0]->attributes()->FormalName
            == $category) && (@preg_match($patternTitle,
            strtolower($title)))) {
                 if(!array_key_exists($file,$such_array)) {
                   $such_array[$file]=$xmlFile;
                   $counter++;
                 }
            }
      } elseif ($patternText != "#()#i") {
            if ($subject->Subject[0]->attributes()->FormalName
            == $category) {
               foreach($body as $weitarr) {
                  foreach($weitarr as $text) {
                      if (@preg_match($patternText,
                      strtolower($text))) {
                          if(!array_key_exists($file,$such_array
                          )) {
                              $such_array[$file]=$xmlFile;
                              $counter++;
                          }
                      }
                  }

               }
            }
      } else {
```

```
            if ($subject->Subject[0]->attributes()->FormalName ==
            $category) {
                if(!array_key_exists($file,$such_array)) {
                     $such_array[$file]=$xmlFile;
                     $counter++;
                }
            }
        }
    }

    if($patternTitle != "#()#i" && $category == "0") {
        if ($patternText != "#()#i") {
            foreach($body as $weitarr) {
                foreach($weitarr as $text) {
                    if (!array_key_exists($file,$such_array) &&
                    (@preg_match($patternTitle,
                    strtolower($title))) &&
                    (@preg_match($patternText,
                    strtolower($text)))) {
                         $such_array[$file]=$xmlFile;
                         $counter++;
                    }
                }
            }
        } else {
             if (@preg_match($patternTitle,
             strtolower($title))) {
                 if(!array_key_exists($file,$such_array)) {
                     $such_array[$file]=$xmlFile;
                     $counter++;
                 }
             }
        }
    }

    if($patternText != "#()#i" && $patternTitle == "#()#i" &&
    $category == "0") {
        foreach($body as $weitarr) {
            foreach($weitarr as $text) {
                 if (@preg_match($patternText,
                 strtolower($text))) {
                     if(!array_key_exists($file,$such_array)) {
                         $such_array[$file]=$xmlFile;
                         $counter++;
                     }
                 }
            }
        }
    }
   } else {
    $inhalt .= $pfad.$file."<br>";
    $fehler++;
   }
}
echo serialize(simplexml2array($such_array));
```

```
 if($fehler>0) {
     mail("bettina.tuller@kurier.at", "Fehler in
     apabasis_simplexml.php", $fehler." Fehler beim Versuch
     folgende Dateien zu laden:<br><br>".$inhalt);
 }
}
?>
```

A.3 dom.php

```
<?
if($_GET["meinvar"] != md5("{Ausdruck}")) {
     echo "nicht berechtigt";
     exit();
}

$pfad = "/apa/basis/";
$filename = $_GET["file"];

if(file_exists($pfad.$filename)) {
    $doc=domDocument::load($pfad.$filename);
    $xpath = new DOMXPath($doc);
    $xpath-
    >registerNamespace("x","http://www.w3.org/1999/xhtml");
    $hl_query = "//HeadLine";
    $headline = $xpath->query($hl_query)->item(0)->nodeValue;

    $shl_query = "//SubHeadLine";
    $subheadline = $xpath->query($shl_query)->item(0)->nodeValue;

    $lc_query = "//Location/Property/@Value";
    $location = $xpath->query($lc_query)->item(0)->nodeValue;

    $b_query = "//DataContent/x:html/x:body";
    $body = $xpath->query($b_query)->item(0)->nodeValue;

    $i_query = "//Instruction/@FormalName";
    $instruction = $xpath->query($i_query)->item(0)->nodeValue;

    $s_query = "//Source/Party";
    $source = $xpath->query($s_query);
    $z = 0;

    foreach ($source as $teile) {
     if($teile->hasAttributes()) {
          foreach ($teile->attributes as $attribute){
               if($z != 0) $sourcen.=", ";
               $sourcen .= $attribute->value;
```

```
            $z++;
           }
 }
}

$d_query = "//DateAndTime";
$date = $xpath->query($d_query)->item(0)->nodeValue;

$c_query = "//Creator/Party";
$creator = $xpath->query($c_query);
$z = 0;

foreach ($creator as $teile) {
 if($teile->hasAttributes()) {
      foreach ($teile->attributes as $attribute){
           if($z != 0) $creatorn.=", ";
           $creatorn .= $attribute->value;
           $z++;
           }
 }
}

$kwl_query = "//KeywordLine";
$keywordline = $xpath->query($kwl_query)->item(0)->nodeValue;

$sc_query = "//SubjectCode/Subject";
$newscode = $xpath->query($sc_query);
$z = 0;

foreach ($newscode as $teile) {
 if($teile->hasAttribute("Vocabulary")) {
      foreach ($teile->attributes as $key => $attribute){
           if($key=="FormalName") {
                if($z != 0) $newscoden.=", ";
                $newscoden .= $attribute->value;
                $z++;
           }
      }
 }
}

$as_query = "//AssociatedWith/@NewsItem";
$associated = $xpath->query($as_query)->item(0)->nodeValue;

$li_query =
'//Metadata/Property[@FormalName="Homepage"]/@Value';
$link = $xpath->query($li_query)->item(0)->nodeValue;

$cl_query = "//CopyrightLine";
$copyline = $xpath->query($cl_query)->item(0)->nodeValue;

$l_query = "//LabelText";
$label = $xpath->query($l_query)->item(0)->nodeValue;

echo $headline.'<bez>'.$subheadline.'<bez>'.$location.'<bez>'
```

```
    .$body.'<bez>'.$sourcen.'<bez>'.$date.'<bez>'.$creatorn.'<bez
    >'
    .$keywordline.'<bez>'.$newscoden.'<bez>'.$associated.'<bez>'
    .$link.'<bez>'.$copyline.'<bez>'.$label.'<bez>'.$instruction;
}  else {
     mail("bettina.tuller@kurier.at", "Fehler in dom.php", "das
     angeforderte File ".$filename." konnte nicht gefunden und
     geladen werden");
}
?>
```

A.4 parsertopic.php

```
<?
include("functions.inc.php");

if($xmltopic = simplexml_load_file("http://www.apa-it.at/NewsML/
IPTC/topicset/topicset.apa-cat.xml")) {

} else {
     $xmltopic = simplexml_load_file("topicset.apa-cat.xml");

     mail("bettina.tuller@kurier.at", "Fehler in
     parsertopic.php", "Fehler beim Laden der APA-Kategorien! Bin
     auf das lokal gespeicherte File ausgewichen, überprüf
     mich!");
}
echo serialize(simplexml2array($xmltopic));
?>
```

A.5 apabasis.php

```
<?
header('Content-Type: text/html; charset=utf-8');
include("functions.inc.php");
$var = md5("{Ausdruck}");
?>

<script type="text/javascript">
function popup (link) {
  fenster = window.open(link, "Popup", "scrollbars=yes");
  fenster.focus();
}
```

```
</script>

<?
if($unser_topic = unserialize(file_get_contents("{Pfad}/
parsertopic.php"))) {

     $subject_arr = $unser_topic["NewsItem"]["TopicSet"]
                    ["Topic"];

     foreach($subject_arr as $arr_im_arr) {
          $category_arr[$arr_im_arr["FormalName"]] =
          $arr_im_arr["Description"][0];
     }

     array_unshift($category_arr, "--select--");
     $categories =1;
} else {
     $categories =0;
     mail("bettina.tuller@kurier.at","Fehler in
     apabasis.php","Fehler beim Übernehmen der APA-Kategorien,
     ich konnte kein File laden, hilf mir!");
}
?>

<img src="img/area_search_small.gif" width="160" height="24"><br>

<table border="0" cellspacing="0" cellpadding="3" width="160">
<form action="" method="POST" name="formapa">
<tr>
<td align="left" valign="top" class="bgredsyssmall">
<img src="img/pixel.gif" width="154" height="7"><br>
<font class="text">Titel<br></font>
<input name="titelsuche" type="CHAR" class="input150" value="<?
echo $_REQUEST["titelsuche"]; ?>">
</td>
</tr>
<tr>
<td align="left" valign="top" class="bgredsyssmall">
<img src="img/pixel.gif" width="154" height="7"><br>
<font class="text">Volltext<br></font>
<input name="textsuche" type="CHAR" class="input150" value="<?
echo $_REQUEST["textsuche"]; ?>">
</td>
</tr>

<? if ($categories) { ?>
<tr>
<td align="left" valign="top" class="bgredsyssmall">
<img src="img/pixel.gif" width="154" height="7"><br>
<font class="text">Kategorien<br></font>
<select name="category" class="input150">
<?
foreach($category_arr as $key => $val) {
     echo "<option ";
     if($_REQUEST["category"]==$key) echo "selected ";
```

```
      echo "value='".$key."'>".$val."</option>";
}
?>
</select>
<img src="img/pixel.gif" width="154" height="7"><br>
<img src="img/ps/bgredsys_spacer_small.gif" width="154"
height="13"><br>
</td>
</tr>
<? } ?>

</table>
<img src="img/area_stop_small.gif" width="160" height="14">

<table border="0" cellspacing="0" cellpadding="3" width="160">
<tr>
<td align="center" valign="top" class="bgredsyssmall">
<img src="img/pixel.gif" width="154" height="7"><br>
<input type="submit" name="npf_search" value="Suchen"
class="search" /> 
<br>
<img src="img/pixel.gif" width="154" height="7"><br>
</td>
</tr>
</form>
</table>

<table border="0" cellpadding="0" cellspacing="0">
<tr>
<td align="left" valign="top">
<img src="img/area_stop_small.gif" width="160" height="14"><br>
</td>
</tr>
</table>

<?

if ($_POST["npf_search"] == "Suchen") {

      $such_array = array();

      $such_title = explode(" ",strip_tags(trim(
                  $_REQUEST["titelsuche"])));

      $anzahl_title = count($such_title) - 1;

      foreach ($such_title as $tkey => $tval) {
            $patternTitle .= strip_tags(strtolower($tval));
            if ($tkey < $anzahl_title) $patternTitle .= "|";
      }

      $such_text = explode(" ",strip_tags(trim(
                  $_REQUEST["textsuche"])));

      $anzahl_text = count($such_text) - 1;
```

```
foreach ($such_text as $tkey => $tval) {
     $patternText .= strip_tags(strtolower($tval));
     if ($tkey < $anzahl_text) $patternText .= "|";
}

if($such_array = unserialize(file_get_contents("{Pfad}/
apabasis_simplexml.php?patterntext=".$patternText."
&patterntitle=".$patternTitle."&category=".
$_REQUEST["category"]."&meinvar=".$var))) {

     if(is_array($such_array)) {
          $anz = count($such_array);
     } else {
          $anz=0;
     }

echo "<table width='160' border='0' cellspacing='0'
cellpadding='0'>
<tr><td width='23' align='left' valign='center'>
<img src='img/i.gif' width='23' height='25'><br>
</td><td width='137' align='left' valign='center'>
<font class='setact'>".$anz." Dokument(e)
gefunden<br></font>
</td></tr></table>
<img src='img/hits_start.gif' width='160'
height='26'><br><center>";

foreach ($such_array as $skey => $sucherg) {

     $filedatum = $sucherg["NewsEnvelope"]["DateAndTime"];
     $urgency = $sucherg["NewsItem"]["NewsManagement"]
          ["Urgency"]["@attributes"]["FormalName"];
     $title = $sucherg["NewsItem"]["NewsComponent"]
          ["ContentItem"]["DataContent"]["html"]["head"]
          ["title"];
     $kurztext = $sucherg["NewsItem"]["NewsComponent"]
          ["NewsLines"]["SubHeadLine"];
     $content = $sucherg["NewsItem"]["NewsComponent"]
          ["ContentItem"]["DataContent"]["html"]["body"];
     $credit = $sucherg["NewsItem"]["NewsComponent"]
          ["AdministrativeMetadata"]["Source"];
     $location = $sucherg["NewsItem"]["NewsComponent"]
          ["DescriptiveMetadata"]["Location"]["Property"]
          ["@attributes"]["Value"];

     echo "<img src='img/pixel.gif' width='1'
     height='7'><br>";
     echo "<a href='router.php?file=".$skey."&meinvar="
     .$var."&?patterntext=".$patternText."&patterntitle="
     .$patternTitle."&category=".$_REQUEST["category"]."'
     style='color:".textfarbe($urgency)."'
     class='textboldbig' onclick='popup(this.href); return
     false'>";
```

```
if($patternTitle!="") {
     echo highlight($patternTitle,$title);
} else {
     echo $title;
}
echo "<br></a>";

if(is_array($credit["Party"])) {
   foreach($credit["Party"] as $key => $value) {
       if($anzcredit==1) {
           $credits .= $value["FormalName"];
       } else {
            if($key==0) {
               $credits .= $value["@attributes"]
               ["FormalName"];
            } else {
                $credits .= ", ".$value["@attributes"]
               ["FormalName"];
            }
       }
   }
} else {
     $credits = $credit;
}

$contents = $location;

foreach($content as $schl => $weitarr) {
     if(is_array($weitarr)) {
        foreach($weitarr as $text) {
            $contents.= "<".$schl.">".$text.
            "</".$schl.">\r\n";
        }
     } else {
        $contents .= "<".$schl.">".$weitarr.
        "</".$schl.">\r\n";
     }
}

echo "<font class='text'>".substr($filedatum,6,2).
".".substr($filedatum,4,2).".".substr($filedatum,0,4).
" - ".substr($filedatum,9,2).":".
substr($filedatum,11,2).":".substr($filedatum,13,2).
"<br></font>
<img src='img/pixel.gif' width='1' height='7'><br>
<a href=\"#\" onclick=\"parent.edit.document.form1.
elements['xmlval_TITLE[0]'].value ='".repl($title)."';
parent.edit.document.form1.elements
['xmlval_SUMMARY[0]'].value = '". repl($kurztext)."';
parent.edit.document.form1.elements
['xmlval_CONTENT[0][C_TEXT][0]'].value = '".
repl($contents)."';
parent.edit.document.form1.elements
['xmlval_CREDIT[0]'].value = '". repl($credits)."';\">
```

```
            <img src='img/erstellen.gif' border='0' title='Artikel
            erstellen'></a>
            <img src='img/area_stop_small.gif' width='160'
            height='14'><br>";

            $zaehler++;
      }
} else {
      echo "Fehler!!!<br><br>Bleiben Sie ruhig, Hilfe ist
      unterwegs!<br>Es wurde ein Mail an den Administrator
      versendet!";

      mail("bettina.tuller@kurier.at", "Fehler in apabasis.php",
      "Fehler beim Laden des such_array aus
      apabasis_simplexml.php. Überprüf mich!");
}
}
?>
```

A.6 router.php

```
<?
header('Content-Type: text/html; charset=utf-8');

if($_GET["meinvar"] != md5("{Ausdruck}")) {
      echo "nicht berechtigt";
      exit();
}

include("functions.inc.php");
$var = md5("{Ausdruck}");
$akt_file = $_GET["file"];

if ($mlneu = file_get_contents("{Pfad}/dom.php?
file=".$akt_file."&meinvar=".$var)) {
      $boom = explode("<bez>",$mlneu);

      foreach ($boom as $key => $teil) {
            $boom[$key] = repl($teil);
      }

      $title = $boom[0];
      $kurztext = $boom[1];
      $location = $boom[2];
      $content = $boom[3];
      $credit = $boom[4];

      if($location!="") {
            $content = $location." - ".$content;
```

```
}

$date = $boom[5];
$creator = $boom[6];
$keyword = $boom[7];
$newscode = $boom[8];
$asso = $boom[9];
$webtipp = $boom[10];
$copyline = $boom[11];
$label = $boom[12];
$instruction = $boom[13];

echo "<html><head>
<title>".$title."</title>
<link rel='stylesheet' type='text/css' href='style.css'>
</head>
<body>
<h2>".$title."</h2>
<h4>".$kurztext."</h4>";

if($instruction!="") {
     $content .= "(".$instruction.")";
}

echo "<p>".$content."</p>";

echo "<p>Datum: ".substr($date,6,2).".".substr($date,4,2).
".".substr($date,0,4)." - ".substr($date,9,2).":".
substr($date,11,2).":".substr($date,13,2)."<br>Creator:
".$creator."<br>\r\nSource: ".$credit."<br><br>\r\nKeywords:
".$keyword."<br>\r\nAPARessort: ".$newscode;

if ($asso!="") {
     $asso = explode(":",$asso);
     $assoc = substr($asso[4],8,7)."-
".substr($asso[4],0,8);

     $edate = substr($asso[4],0,8);
     $jahr = substr($edate,0,4);
     $monat = substr($edate,4,2);
     $tag = substr($edate,6,2);

     $tmstmp = mktime(0,0,0,$monat,$tag,$jahr);
     $max = time()-(60*60*24*1);

     if($tmstmp>=$max) {
          $associatedfile = $assoc.".xml";
          $headlineasso = file_get_contents("{Pfad}/
          dom.php?file=".$associatedfile."
          &meinvar=".$var);
          $booma = explode("<bez>",$headlineasso);
     }
     echo "<br><br>\r\nVerwandter Artikel: <a
     href='{Pfad}/router.php?file=".$associatedfile."
     &meinvar=".$var."' target='_blank'>".$booma[0]."</a>";
```

```
        }

        if($webtipp!="") {
              echo "<br><br>\r\nWebtipp: <a href='".$webtipp."'
              target='_blank'>".$webtipp."</a>";
        }

        echo "<br><br>\r\nCopyright: ".$copyline."<br>\r\nAPA-Nr.:
        ".$label."</p></body>\r\n</html>";
} else {
        echo "Fehler!!!<br><br>Bleiben Sie ruhig, Hilfe ist
        unterwegs!<br>Es wurde ein Mail an den Administrator
        versendet!";

        mail("bettina.tuller@kurier.at", "Fehler in router.php",
        "Fehler beim Laden der Inhalte aus dom.php. Überprüf
        mich!");
}
?>
```

Anhang B: NewsML

B.1 NewsML von Reuters

```
<?xml version="1.0" encoding="UTF-8"?>
<NewsML Duid="IIMNSR00001_2006-06-20_13-12-57_L20884816_NEWSML">
    <Catalog Href="../ref/catalog-reuters-master_catalog_1.xml"/>
    <NewsEnvelope>
        <DateAndTime>20060620T131257+0000</DateAndTime>
        <NewsService FormalName="RTR_TNS"/>
        <NewsProduct FormalName="TXT"/>
        <Priority FormalName="3"/>
    </NewsEnvelope>
    <NewsItem Duid="IIMNSR00001_2006-06-20_13-12-
    57_L20884816_NEWSITEM">
        <Identification>
            <NewsIdentifier>
                <ProviderId>reuters.com</ProviderId>
                <DateId>20060620</DateId>
                <NewsItemId>IIMNSR00001_2006-06-20_13-12-
                57_L20884816</NewsItemId>
                <RevisionId Update="N"
                PreviousRevision="0">1</RevisionId>
                <PublicIdentifier>urn:newsml:reuters.com:20060620
                :IIMNSR00001_2006-06-20_13-12-
                57_L20884816:1</PublicIdentifier>
            </NewsIdentifier>
            <DateLabel>2006-06-20 13:12:57 GMT
            (Reuters)</DateLabel>
        </Identification>
        <NewsManagement>
            <NewsItemType FormalName="News"/>
            <FirstCreated>20060620T131257+0000</FirstCreated>
            <ThisRevisionCreated>20060620T131257+0000
            </ThisRevisionCreated>
            <Status FormalName="Usable"/>
            <Urgency FormalName="3"/>
        </NewsManagement>
        <NewsComponent EquivalentsList="no" Essential="no"
        Duid="IIMNSR00001_2006-06-20_13-12-57_L20884816_MAIN_NC"
        xml:lang="de">
        <TopicSet FormalName="HighImportance">
           <Topic Duid="ts_1">
                <TopicType FormalName="SupplementalCategory"/>
                <FormalName Scheme="N2000">GEG</FormalName>
                <Property FormalName="WhyPresent"
                Value="Classifier"/>
           </Topic>
```

```
    <Topic Duid="ts_2">
         <TopicType FormalName="SupplementalCategory"/>
         <FormalName Scheme="N2000">DE</FormalName>
         <Property FormalName="WhyPresent"
         Value="Classifier"/>
    </Topic>
    <Topic Duid="ts_3">
         <TopicType FormalName="SupplementalCategory"/>
         <FormalName Scheme="N2000">AT</FormalName>
         <Property FormalName="WhyPresent"
         Value="Classifier"/>
    </Topic>
    <Topic Duid="ts_4">
         <TopicType FormalName="SupplementalCategory"/>
         <FormalName Scheme="N2000">WEU</FormalName>
         <Property FormalName="WhyPresent"
         Value="Classifier"/>
    </Topic>
    <Topic Duid="ts_5">
         <TopicType FormalName="SupplementalCategory"/>
         <FormalName Scheme="N2000">EUROPE</FormalName>
         <Property FormalName="WhyPresent"
         Value="Classifier"/>
    </Topic>
    <Topic Duid="ts_6">
         <TopicType FormalName="SupplementalCategory"/>
         <FormalName Scheme="N2000">EU</FormalName>
         <Property FormalName="WhyPresent"
         Value="Classifier"/>
    </Topic>
    <Topic Duid="ts_7">
         <TopicType FormalName="SupplementalCategory"/>
         <FormalName Scheme="N2000">FR</FormalName>
         <Property FormalName="WhyPresent"
         Value="Classifier"/>
    </Topic>
</TopicSet>
<Role FormalName="Main"/>
    <AdministrativeMetadata>
        <FileName>MDF87664.II2</FileName>
        <Provider>
            <Party FormalName="Reuters"/>
        </Provider>
        <Source>
            <Party FormalName=""/>
        </Source>
        <Property FormalName="SourceFeed" Value="MDN"/>
        <Property FormalName="IDSPublisher"
        Value="http://www.reuters.com/ids"/>
    </AdministrativeMetadata>
    <NewsComponent EquivalentsList="no" Essential="no"
    Duid="IIMNSR00001_2006-06-20_13-12-
    57_L20884816_MAIN_TEXT_NC" xml:lang="de">
        <Role FormalName="Main Text"/>
        <NewsLines>
```

```
        <HeadLine>Schüssel - EU-weites Referendum
         einziger Weg zur Verfassung</HeadLine>
        <ByLine> </ByLine>
        <DateLine>June 20, 2006</DateLine>
        <CreditLine>Reuters</CreditLine>
        <CopyrightLine>(c) Reuters 2006. All rights
         reserved. Republication or redistribution of
         Reuters content, including by caching,
         framing or similar means, is expressly
         prohibited without the prior written consent
         of Reuters. Reuters and the Reuters sphere
         logo are registered trademarks and trademarks
         of the Reuters group of companies around the
         world.</CopyrightLine>
        <SlugLine>EU/VERFASSUNG/SCHUESSEL</SlugLine>
        <KeywordLine></KeywordLine>
        <NewsLine>
            <NewsLineType FormalName="Caption"/>
            <NewsLineText>EU/VERFASSUNG/SCHUESSEL:Sch
            üssel - EU-weites Referendum einziger Weg
            zur Verfassung</NewsLineText>
        </NewsLine>
    </NewsLines>
    <DescriptiveMetadata>
        <Language FormalName="de"/>
        <OfInterestTo FormalName="GEA"/>
         <OfInterestTo FormalName="SWI"/>
         <OfInterestTo FormalName="OE"/>
         <OfInterestTo FormalName="GERT"/>
         <OfInterestTo FormalName="DNP"/>
        <TopicOccurrence Importance="High"
         Topic="ts_1"/>
         <TopicOccurrence Importance="High"
         Topic="ts_2"/>
         <TopicOccurrence Importance="High"
         Topic="ts_3"/>
         <TopicOccurrence Importance="High"
         Topic="ts_4"/>
         <TopicOccurrence Importance="High"
         Topic="ts_5"/>
         <TopicOccurrence Importance="High"
         Topic="ts_6"/>
         <TopicOccurrence Importance="High"
         Topic="ts_7"/>
    </DescriptiveMetadata>
    <ContentItem Duid="IIMNSR00001_2006-06-20_13-12-
     57_L20884816_MAIN1_TEXT_CI">
        <MediaType FormalName="Text"/>
        <Format FormalName="XHTML"/>
        <Characteristics>
            <Property FormalName="ContentID"
             Value="urn:newsml:reuters.com:20060620:II
             MNSR00001_2006-06-20_13-12-
             57_L20884816_TXT:1"/>
```

```
                        <Property
                         FormalName="ContentCreationDateAndTime"
                         Value="20060620T131257+0000"/>
                        <Property FormalName="USN"
                         Value="L20884816"/>
                        <Property FormalName="Creator"
                         Value="NewsSuite v3"/>
                    </Characteristics>
                    <DataContent>
                        <html
                         xmlns="http://www.w3.org/1999/xhtml">
                            <head>
                                <title/>
                            </head>
                            <body>
                                <p>    Brüssel, 20. Jun (Reuters)
                                - Österreichs Bundeskanzler
                                Wolfgang Schüssel sieht ein EU-
                                weites Referendum über die
                                Europäische Verfassung parallel
                                zur Europawahl als einzigen, aber
                                unwahrscheinlichen Weg aus der
                                Verfassungskrise.
                                </p>
                                <p>    Nach seiner Abschiedsrede
                                vor dem EU-Parlament verteidigte
                                der scheidende EU-Ratspräsident
                                seinen Vorschlag für eine EU-weite
                                Volksabstimmung. Wenn die EU-
                                Bürger bei der nächsten Wahl zum
                                Europäischen Parlament 2009 auch
                                zur Verfassung befragt würden,
                                gebe es die Chance auf ein echtes
                                Votum zu dem Text, sagte er am
                                Dienstag in Brüssel. Schüssel
                                sprach sich für eine doppelte
                                Mehrheit aus Bevölkerung und
                                Staaten aus. Damit sind in der EU
                                Abstimmungsmodelle gemeint, bei
                                denen einzelne Länder auch
                                überstimmt werden können.
                                </p>
                                <p>    cal/phi </p>
                            </body>
                        </html>
                    </DataContent>
                </ContentItem>
            </NewsComponent>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

B.2 NewsML von AFP (Agence France Presse)

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<NewsML>
      <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
      <NewsEnvelope>
            <DateAndTime>20030319T000031Z</DateAndTime>
            <NewsService FormalName="FRSGL"/>
            <Priority FormalName="3"/>
      </NewsEnvelope>
      <NewsItem>
            <Identification>
                  <NewsIdentifier>
                        <ProviderId>afp.com</ProviderId>
                        <DateId>20030319</DateId>
                        <NewsItemId>000031-TX-LNH41</NewsItemId>
                        <RevisionId PreviousRevision="0"
                        Update="N">1</RevisionId>
                        <PublicIdentifier>urn:newsml:afp.com:200303
                        19:000031-TX-LNH41:1</PublicIdentifier>
                  </NewsIdentifier>
                  <NameLabel>Irak-Onu</NameLabel>
                  <Label>
                        <LabelType FormalName="SequenceNumber"/>
                        <LabelText>21</LabelText>
                  </Label>
            </Identification>
            <NewsManagement>
                  <NewsItemType FormalName="News"/>
                  <FirstCreated>20030319T000031Z</FirstCreated>
                  <ThisRevisionCreated>20030319T000031Z
                  </ThisRevisionCreated>
                  <Status FormalName="Usable"/>
            </NewsManagement>
            <NewsComponent>
                  <NewsLines>
                        <HeadLine>Bagdad rejette l'ultimatum, le
                        Parlement britannique autorise la force
                        (SYNTHESE ACTUALISEE A 00H00
                        GMT)</HeadLine>
                        <DateLine>BAGDAD, 18 mars (AFP) -</DateLine>
                        <CopyrightLine>© 2002 AFP</CopyrightLine>
                        <SlugLine>Irak-Onu PREV-SYNTH </SlugLine>
                  </NewsLines>
                  <AdministrativeMetadata>
                        <Provider>
                              <Party FormalName="AFP"/>
                        </Provider>
                        <Creator>
                              <Party FormalName="BUR"/>
                        </Creator>
                  </AdministrativeMetadata>
```

```
                  <DescriptiveMetadata>
                        <Language FormalName="fr-FR"/>
                        <SubjectCode>
                              <Subject
                              Vocabulary="urn:newsml:afp.com:2001100
                              1:AFPCatCodes:1" FormalName="I"/>
                        </SubjectCode>
                        <SubjectCode>
                              <Subject
                              Vocabulary="urn:newsml:afp.com:2001100
                              1:AFPCatCodes:1" FormalName="P"/>
                        </SubjectCode>
                        <SubjectCode>
                             <Subject FormalName="11000000"/>
                        </SubjectCode>
                        <Property FormalName="Location">
                              <Property FormalName="Country"
                              Value="IRQ"/>
                              <Property FormalName="City"
                              Value="BAGDA"/>
                        </Property>
                  </DescriptiveMetadata>
                  <ContentItem>
                        <MediaType FormalName="Text"/>
                        <Format FormalName="bcNITF2.5"/>
                        <Characteristics>
                              <Property FormalName="WordCount"
                              Value="823"/>
                        </Characteristics>
                        <DataContent>
                              <p>L'Irak a rejeté mardi l'ultimatum
                              de George W. Bush exigeant le départ
                              d'ici jeudi 01h00 GMT du président
                              Saddam Hussein, tandis qu'à Londres,
                              le Parlement britannique adoptait une
                              motion du gouvernement autorisant
                              l'usage de la force contre le régime
                              du président irakien.</p>
                              <p>Cette motion a été approuvée par
                              412 des 659 députés de la chambre des
                              Communes, mais 149 élus ont voté
                              contre. La Grande-Bretagne a mobilisé
                              45.000 soldats pour une guerre contre
                              l'Irak.</p>
                             <p>bur-jr/prh </p>
                        </DataContent>
                  </ContentItem>
            </NewsComponent>
      </NewsItem>
</NewsML>
```

B.3 „NewsML by APA“

```
<?xml version="1.0" encoding="windows-1252"?>
<NewsML>
 <Catalog Href="http://www.apa-it.at/NewsML/catalog/
                catalog.APAMasterCatalog.xml"/>
 <NewsEnvelope>
    <DateAndTime>20060604T035642+0200</DateAndTime>
    <NewsService FormalName="Agencies"/>
    <NewsProduct FormalName="APA Basisdienst"/>
  </NewsEnvelope>
   <NewsItem>
     <Identification>
       <NewsIdentifier>
         <ProviderId>apa.at</ProviderId>
         <DateId>20060604</DateId>
         <NewsItemId>20060604APA0017</NewsItemId>
            <RevisionId PreviousRevision="0"
            Update="N">1</RevisionId>
            <PublicIdentifier>urn:newsml:apa.at:20060604:20060604A
            PA0017:1</PublicIdentifier>
       </NewsIdentifier>
       <Label>
         <LabelType FormalName="APAIntraId"/>
         <LabelText>APA0017</LabelText>
       </Label>
        <Label>
         <LabelType FormalName="APAIntraDate"/>
         <LabelText>2006-06-04</LabelText>
       </Label>
     </Identification>
     <NewsManagement>
       <NewsItemType FormalName="News"/>
       <FirstCreated>20060604T035642+0200</FirstCreated>
      <ThisRevisionCreated>20060604T035642+0200
      </ThisRevisionCreated>
       <Status FormalName="Usable"/>
       <Urgency FormalName="5"/>
       <Instruction FormalName="fortsetzung-moeglich"/>
     </NewsManagement>
     <NewsComponent>
       <Role FormalName="Hauptinhalt"/>
       <NewsLines>
           <HeadLine>China: Militärflugzeug mit 40 Insassen
           abgestürzt - Agentur</HeadLine>
           <SubHeadLine>Utl.: Im Osten des Landes - Beamter: Fünf
           Leichen geborgen </SubHeadLine>
           <CopyrightLine>Copyright, Austria Presse Agentur, Wien,
           Österreich, Alle Rechte vorbehalten</CopyrightLine>
           <KeywordLine>Unfälle/Verkehr/Luftfahrt/China
           </KeywordLine>
       </NewsLines>
```

```
      <AdministrativeMetadata>
        <Provider><Party FormalName="APA"/></Provider>
        <Creator><Party FormalName="vos"/></Creator>
        <Source>
          <Party FormalName="APA"/>
          <Party FormalName="Reuters"/>
          <Party FormalName="AP"/>
        </Source>
      </AdministrativeMetadata>
      <DescriptiveMetadata>
         <Language FormalName="de-AT"
         Vocabulary="http://www.ietf.org/rfc/rfc3066.txt"/>
         <SubjectCode>
            <Subject FormalName="CA"
            Vocabulary="urn:newsml:apa.at:20040101:APAcat:1"/>
            <Subject FormalName="AA"
            Vocabulary="urn:newsml:apa.at:20040101:APAcat:1"/>
            <Subject FormalName="08000000"/>
            <Subject FormalName="11000000"/>
         </SubjectCode>
         <Location HowPresent="Origin">
           <Property FormalName="City" Value="Peking"/>
         </Location>
         <Property FormalName="APAstwort" Value="Unfälle"/>
         <Property FormalName="APAstwort" Value="Verkehr"/>
         <Property FormalName="APAstwort" Value="Luftfahrt"/>
         <Property FormalName="APAstwort" Value="China"/>
      </DescriptiveMetadata>
       <ContentItem>
         <MediaType FormalName="Text"/>
         <Format FormalName="XHTML"/>
         <DataContent>
           <html xmlns="http://www.w3.org/1999/xhtml">
             <head>
                 <title>China: Militärflugzeug mit 40 Insassen
                 abgestürzt - Agentur</title>
             </head>
             <body>
                 <p>In China ist einer Nachrichtenagentur zufolge
                 am Samstag ein Transportflugzeug des Militärs mit
                 mindestens 40 Menschen an Bord abgestürzt.
                 Experten aus Peking leiteten den Bergungseinsatz
                 in der Provinz Anhui, meldete die Agentur Xinhua
                 am Sonntag unter Berufung auf Militärkreise. Fünf
                 Leichen seien an der Absturzstelle im Bezirk
                 Guangde bereits geborgen worden, sagte ein
                 Beamter der Regionalbehörde. Guangde liegt 200
                 Kilometer südwestlich von Shanghai.</p>
             </body>
           </html>
         </DataContent>
       </ContentItem>
     </NewsComponent>
   </NewsItem>
</NewsML>
```