DaCaPo-intelligente Inhaltserfassung von Zeitungsartikeln

Dr. Wolfgang Schade
Gesellschaft zur Förderung angewandter Informatik e.V. (GFaI)
Forschungsbereich Dokumentenmanagement
Volmerstr. 3, 12489 Berlin
Tel.: 030 814 563 470, Fax: 030 814 563 302
E-Mail: schade@gfai.de, Internet: www.gfai.de

Das Erfassungssystem DaCaPo, eine Software-Lösung der GFaI zur **intelligenten Inhaltserfassung von Zeitungsartikeln,** ist eine im Rahmen der INNOKOM-Ost-Förderung (BMWi) weiterentwickelte Version der auch auf der EVA 2011 Berlin vorgestellten Version und hat folgende, auch separat nutzbare, Pakete:

1. ***Preparator***
   Ausrichtung der Images
   Beschneidung des Randes
   Bereitstellung eines Binärbildes

2. ***Analysator***
   Automatische Separierung von:
   Zeitungsköpfen
   (erweiterten) Artikelüberschriften
   Artikeltextblöcken
   Abbildungen
   Bildunterschriften
   Stempeln
   Die Textfelder werden an eine kommerzielle OCR übergeben.

3. ***Evaluator***
   Zur automatischen intelligenten Auswertung der Text- und Stempelfelder:
   Bestimmung des Zeitungsnamens und des Erscheinungsdatums
   Bestimmung der Artikelart (Interview, Roman, Traueranzeige, Kurzbiografie)
   Autor des Artikels
   Abbildungsquelle(Fotograf, Zeichner, Archiv, Pressedienst)
   Textblockreihenfolge unter Zuhilfenahme der OCR-Ergebnisse

4. ***Korrektor***
   Oberfläche zur Kontrolle und interaktiven Korrektur der Ergebnisse
   Anzeige des Farb-Images mit Zoomfunktion
   Korrektur/Löschung/Neuanlage von
   - Bereichen
   - Texten (auch separate OCR-Erkennung möglich)
   - Textblockreihenfolgen
   - Autorennamen und Abbildungsquellen

   Ablage der Ergebnisse in XML

5. ***MySQL-Datenbank***
   Die in der MySQL-Datenbank abgelegten Resultate lassen sich sowohl hausintern wie auch für Internet-Präsentationen nutzen.

6. Das System kann sowohl in einer Stand-alone-Variante als auch als Client-Server-System installiert werden.