DaCaPo: Ein System zur strukturierten Inhaltserfassung von Zeitungen

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 ist eine Client-Server-Anwendung zur teilautomatisierten Inhaltserfassung von Dokumenten.

Mit dem Programmsystem zur interaktiven intelligenten Inhaltserfassung von Zeitungsartikeln aus deren Seiten-Images kann ausgeführt werden:

1. Anlage des Ordners:
   Themengebiet und Signatur des entsprechenden Ordners

2. Artikelerfassung:
   Namenserfassung (Vorname, Nachname, Vorsatz (Graf..), Nachsatz (von..), Titel (Dr.))
   Sprachklassifizierung des Artikels (deutsch, polnisch, tschechisch, ...)
   Anzeige des Scan-Images mit Zoomfunktion
   Textausrichtung (bei schräg aufgeklebten Artikeln)
   Textbereichsseparierung
   Abbildungsseparierung
   Abbildungsbeschreibung (Karikatur, Foto, Skizze)
   Zuordnung von Bildunterschriften
   Bildinhaltsbeschreibung, falls notwendig
   Erfassung von Autoren und Fotografen
   Erfassung der Artikelüberschrift(en)
   Artikelklassifizierung (Anzeige, Gedicht, Reportage, Interview, Roman)
   Kennzeichnung, ob Artikel und/oder Abbildung(en) freigegeben werden können
   gesonderte Erfassung des Zeitungsnamens
   gesonderte Erfassung des Erscheinungsdatums

3. Eintragung der Ergebnisse in eine MySQL-Datenbank

4. **Mit dem auf einem Server laufenden entwickelten Tool werden Überschrifts-, Text-Bild-, Stempel- und Bildunterschriftsbereiche separiert und die von einer kommerziellen OCR gelieferten Ergebnisse in die MySQL-Datenbank eingetragen. Neu ist, dass der Artikel in der Datenbank automatisch angelegt wird und dabei die Textblöcke in der Lesereihenfolge angeordnet werden.**

5. Das Interface bietet außerdem die Möglichkeit, Ergänzungen und Korrekturen hinzuzufügen, wie z. B persönliche Daten der Autoren und PND, und die durch die eingebundene OCR gelieferten Ergebnisse zu kontrollieren bzw. zu korrigieren

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