Neue Verfahren der Inhaltserschließung

Berliner Herbsttreffen zur Museumsdokumentation 2002

Dipl.-Inf. wiss. Manfred Hauer M.A.

AGI
INFORMATION
MANAGEMENT
CONSULTANTS

Neustadt an der Weinstrasse
Germany

1. Beispiel Vorarlberger Landesbibliothek

Begründung für Investition

60 - 70 % der Benutzer suchen nach inhaltlichen Kriterien, Sachbegriffen

Nutzer wenden engst möglichen Suchbegriff an -
Buchtitel sind meist zu allgemein, viele relevante Titel werden nicht gefunden

Benutzer recherchierend zunehmend vom Internet her -
d.h. kein räumliche Nähe zum Medium

Bibliothekskapital "Information" liegt brach - geringer Wirkungsgrad

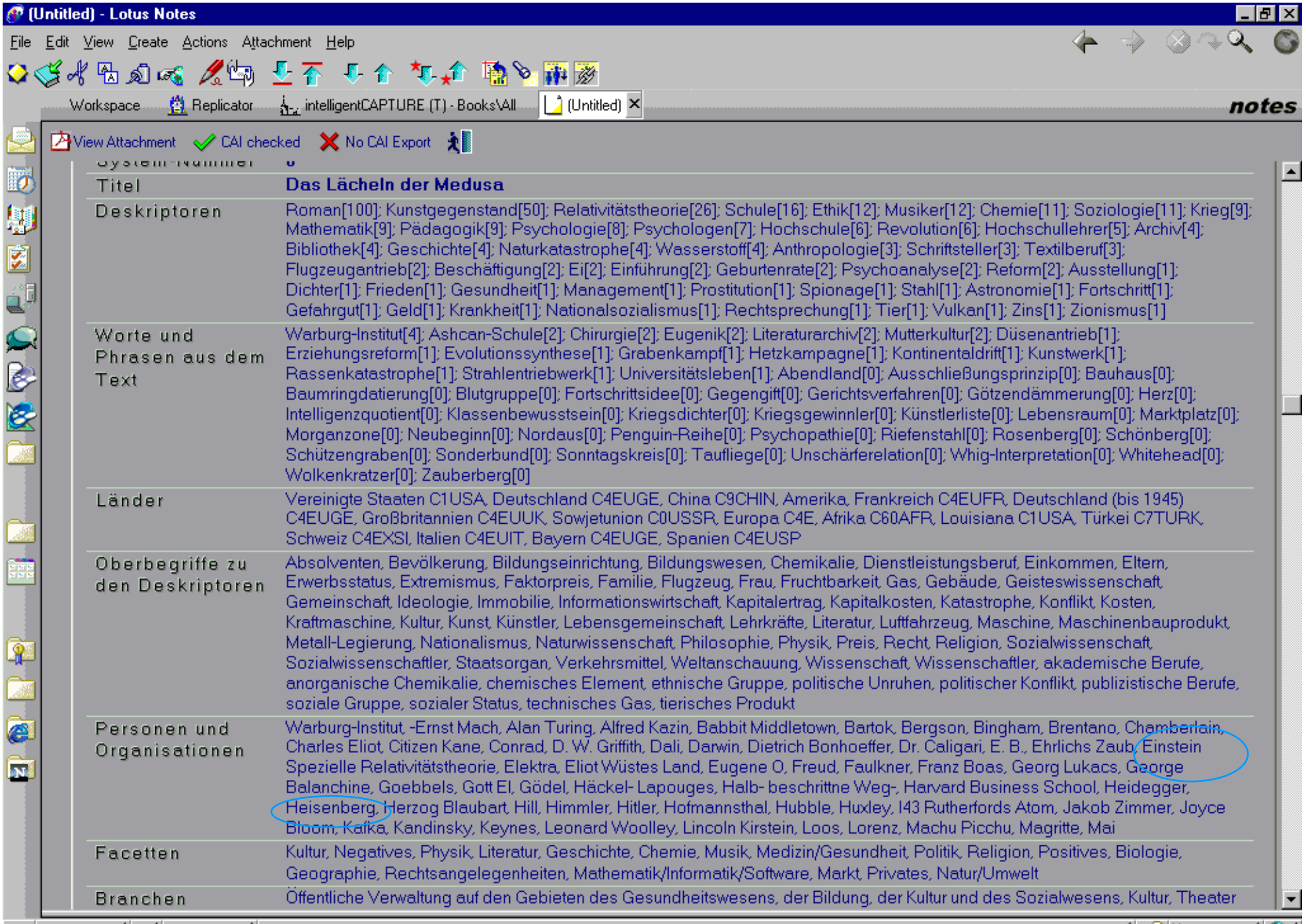
ALEPH im Web

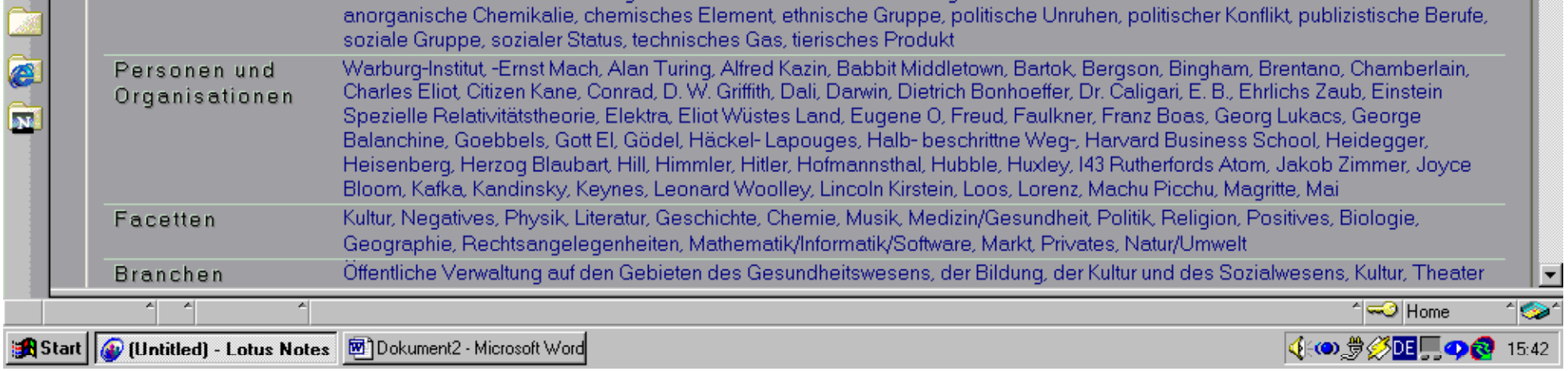
1.1 Demo: Beispiel "Einstein, Heisenberg"

Suche in ALEPH - Feld Inhaltsverzeichnis

http://avlrpk07.br.vlr.gv.at/ALEPH/THIQUK1MNF8P5FBL8626B55HHKTSJGKCKBN79FGNYU4RDASQBV-00348/file/start-0 

Diese Daten sind in ALEPH importiert - ohne die Nummern = Termgewichte, darin recherchiert der Nutzer wirklich, nicht im PDF


(Untitled) - Lotus Notes
File Edit View Create Actions Attachment Help
Workspace
Replicator
intelligentCAPTURE (T) - Books\All
(Untitled)
notes
View Attachment
CAI checked
No CAI Export
Titel
Das Lächeln der Medusa
Deskriptoren
Roman[100]; Kunstgegenstand[50]; Relativitätstheorie[26]; Schule[16]; Ethik[12]; Musiker[12]; Chemie[11]; Soziologie[11]; Krieg[9]; Mathematik[9]; Pädagogik[9]; Psychologie[8]; Psychologen[7]; Hochschule[6]; Revolution[6]; Hochschullehrer[5]; Archiv[4]; Bibliothek[4]; Geschichte[4]; Naturkatastrophe[4]; Wasserstoff[4]; Anthropologie[3]; Schriftsteller[3]; Textilberuf[3]; Flugzeugantrieb[2]; Beschäftigung[2]; Ei[2]; Einführung[2]; Geburtenrate[2]; Psychoanalyse[2]; Reform[2]; Ausstellung[1]; Dichter[1]; Frieden[1]; Gesundheit[1]; Management[1]; Prostitution[1]; Spionage[1]; Stahl[1]; Astronomie[1]; Fortschritt[1]; Gefahrgut[1]; Geld[1]; Krankheit[1]; Nationalsozialismus[1]; Rechtsprechung[1]; Tier[1]; Vulkan[1]; Zins[1]; Zionismus[1]
Worte und Phrasen aus dem Text
Warburg-Institut[4]; Ashcan-Schule[2]; Chirurgie[2]; Eugenik[2]; Literaturarchiv[2]; Mutterkultur[2]; Düsenantrieb[1]; Erziehungsreform[1]; Evolutionssynthese[1]; Grabenkampf[1]; Hetzkampagne[1]; Kontinentaldrift[1]; Kunstwerk[1]; Rassenkatastrophe[1]; Strahlentriebwerk[1]; Universitätsleben[1]; Abendland[0]; Ausschließungsprinzip[0]; Bauhaus[0]; Baumringdatierung[0]; Blutgruppe[0]; Fortschrittsidee[0]; Gegengift[0]; Gerichtsverfahren[0]; Götzendämmerung[0]; Herz[0]; Intelligenzquotient[0]; Klassenbewusstsein[0]; Kriegsdichter[0]; Kriegsgewinnler[0]; Künstlerliste[0]; Lebensraum[0]; Marktplatz[0]; Morganzone[0]; Neubeginn[0]; Nordaus[0]; Penguin-Reihe[0]; Psychopathie[0]; Riefenstahl[0]; Rosenberg[0]; Schönberg[0]; Schützengraben[0]; Sonderbund[0]; Sonntagskreis[0]; Taufliege[0]; Unschärferelation[0]; Whig-Interpretation[0]; Whitehead[0]; Wolkenkratzer[0]; Zauberberg[0]
Länder
Vereinigte Staaten C1USA, Deutschland C4EUGE, China C9CHIN, Amerika, Frankreich C4EUFR, Deutschland (bis 1945) C4EUGE, Großbritannien C4EUUK, Sowjetunion C0USSR, Europa C4E, Afrika C60AFR, Louisiana C1USA, Türkei C7TURK, Schweiz C4EXSI, Italien C4EUIT, Bayern C4EUGE, Spanien C4EUSP
Oberbegriffe zu den Deskriptoren
Absolventen, Bevölkerung, Bildungseinrichtung, Bildungswesen, Chemikalie, Dienstleistungsberuf, Einkommen, Eltern, Erwerbsstatus, Extremismus, Faktorpreis, Familie, Flugzeug, Frau, Fruchtbarkeit, Gas, Gebäude, Geisteswissenschaft, Gemeinschaft, Ideologie, Immobilie, Informationswirtschaft, Kapitalertrag, Kapitalkosten, Katastrophe, Konflikt, Kosten, Kraftmaschine, Kultur, Kunst, Künstler, Lebensgemeinschaft, Lehrkräfte, Literatur, Luftfahrzeug, Maschine, Maschinenbauprodukt, Metall-Legierung, Nationalismus, Naturwissenschaft, Philosophie, Physik, Preis, Recht, Religion, Sozialwissenschaft, Sozialwissenschaftler, Staatsorgan, Verkehrsmittel, Weltanschauung, Wissenschaft, Wissenschaftler, akademische Berufe, anorganische Chemikalie, chemisches Element, ethnische Gruppe, politische Unruhen, politischer Konflikt, publizistische Berufe, soziale Gruppe, sozialer Status, technisches Gas, tierisches Produkt
Personen und Organisationen
Warburg-Institut, -Ernst Mach, Alan Turing, Alfred Kazin, Babbit Middletown, Bartok, Bergson, Bingham, Brentano, Chamberlain, Charles Eliot, Citizen Kane, Conrad, D. W. Griffith, Dali, Darwin, Dietrich Bonhoeffer, Dr. Caligari, E. B., Ehrlichs Zaub, Einstein Spezielle Relativitätstheorie, Elektra, Eliot Wüstes Land, Eugene O, Freud, Faulkner, Franz Boas, Georg Lukacs, George Balanchine, Goebbels, Gott El, Gödel, Häckel- Lapouges, Halb- beschrittne Weg-, Harvard Business School, Heidegger, Heisenberg, Herzog Blaubart, Hill, Himmler, Hitler, Hofmannsthal, Hubble, Huxley, I43 Rutherfords Atom, Jakob Zimmer, Joyce Bloom, Kafka, Kandinsky, Keynes, Leonard Woolley, Lincoln Kirstein, Loos, Lorenz, Machu Picchu, Magritte, Mai
Facetten
Kultur, Negatives, Physik, Literatur, Geschichte, Chemie, Musik, Medizin/Gesundheit, Politik, Religion, Positives, Biologie, Geographie, Rechtsangelegenheiten, Mathematik/Informatik/Software, Markt, Privates, Natur/Umwelt
Branchen
Öffentliche Verwaltung auf den Gebieten des Gesundheitswesens, der Bildung, der Kultur und des Sozialwesens, Kultur, Theater


anorganische Chemikalie, chemisches Element, ethnische Gruppe, politische Unruhen, politischer Konflikt, publizistische Berufe, soziale Gruppe, sozialer Status, technisches Gas, tierisches Produkt
Personen und Organisationen
Warburg-Institut, -Ernst Mach, Alan Turing, Alfred Kazin, Babbit Middletown, Bartok, Bergson, Bingham, Brentano, Chamberlain, Charles Eliot, Citizen Kane, Conrad, D. W. Griffith, Dali, Darwin, Dietrich Bonhoeffer, Dr. Caligari, E. B., Ehrlichs Zaub, Einstein Spezielle Relativitätstheorie, Elektra, Eliot Wüstes Land, Eugene O, Freud, Faulkner, Franz Boas, Georg Lukacs, George Balanchine, Goebbels, Gott El, Gödel, Häckel- Lapouges, Halb- beschrittne Weg-, Harvard Business School, Heidegger, Heisenberg, Herzog Blaubart, Hill, Himmler, Hitler, Hofmannsthal, Hubble, Huxley, I43 Rutherfords Atom, Jakob Zimmer, Joyce Bloom, Kafka, Kandinsky, Keynes, Leonard Woolley, Lincoln Kirstein, Loos, Lorenz, Machu Picchu, Magritte, Mai
Facetten
Kultur, Negatives, Physik, Literatur, Geschichte, Chemie, Musik, Medizin/Gesundheit, Politik, Religion, Positives, Biologie, Geographie, Rechtsangelegenheiten, Mathematik/Informatik/Software, Markt, Privates, Natur/Umwelt
Branchen
Öffentliche Verwaltung auf den Gebieten des Gesundheitswesens, der Bildung, der Kultur und des Sozialwesens, Kultur, Theater
Home
Start
(Untitled) - Lotus Notes
Dokument2 - Microsoft Word
DE
15:42

2. Projekt intelligentCAPTURE beim BMBF

Ziele

1. SWD in maschinelle Indexierung integrieren
2. DIPF-Thesaurus in m. I. integrieren und dynamisch weiterentwickeln
3. Praxis-Eignung weiter erforschen
4. Software-Entwicklung weiter vorantreiben

Partner

Bibliotheksservice-Zentren	BSZ Konstanz GBV Göttingen
Bibliotheken & Dokumentation	TIB Hannover Universitätsbibliothek Konstanz Vorarlberger Landesbibliothek Deutsche Institut für internationale pädagogische Forschung (DIPF)

Software &
Language Engineering

AGI - Information Management Consultants
IAI - Saarbrücken

3. Nutzen für Informationssuchende

1. Hochspezifische Suchterme sind möglich und erfolgreich im OPAC

2. Zusätzliche Stringsuche im PDF-Inhaltsverzeichnis möglich

3. Nur relevante Terme aus dem Text werden indexiert - kein Volltext-Overhead

4. Reduktion auf Grundformen, sinnvolle Mehrwortbegriffe - Vereinfachung der Suche

5. Inhaltsverzeichnis kann direkt eingesehen werden - schnellere Auswahl relevanter Titel

6. Deutlich höherer Recall - alle relevanten Artikel werden gefunden

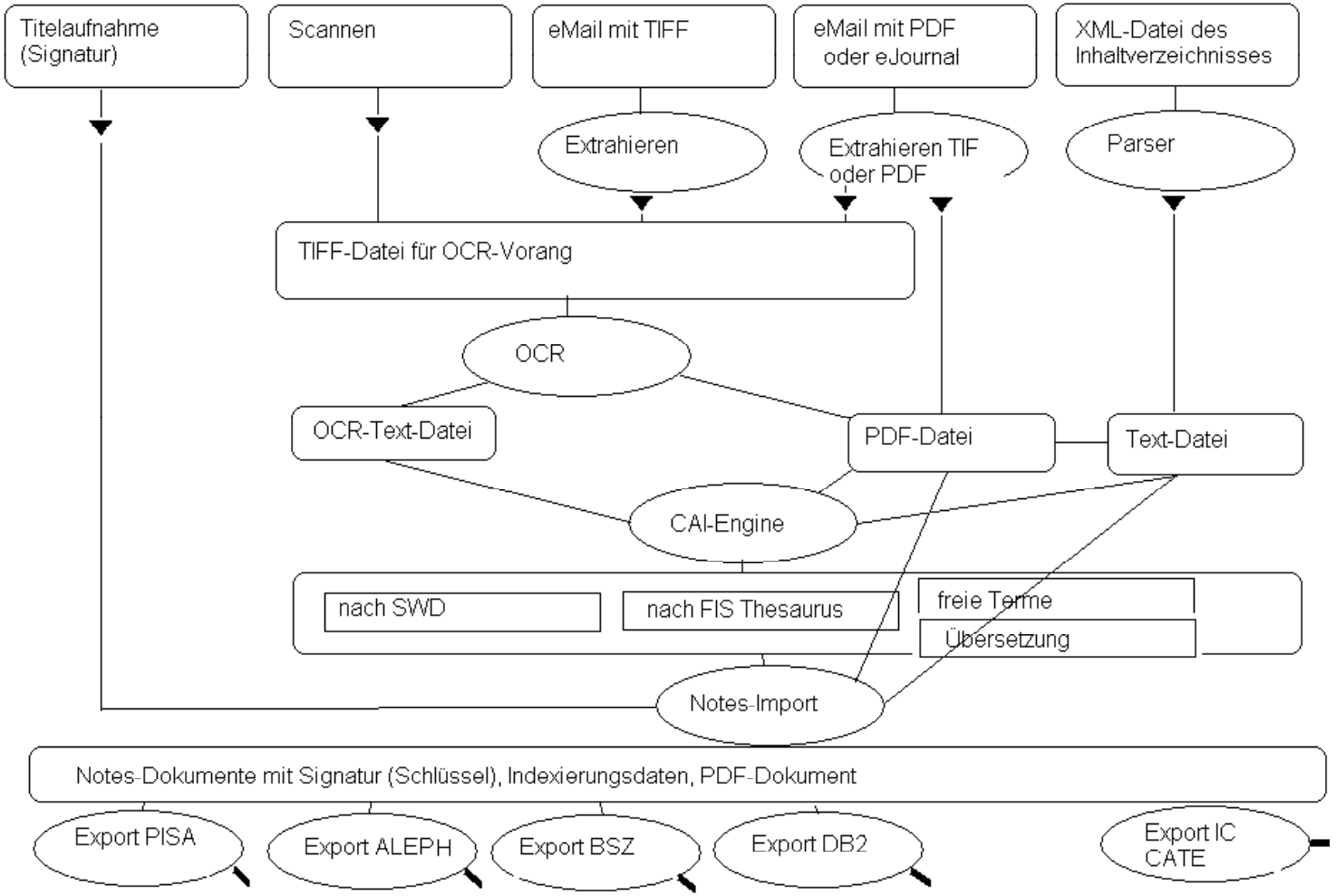
7. Deutlich höhere Precission - nur Dokumente, in denen hochspezifische Terme eine wichtige Rolle spielen werden gefunden. Nutzer kann durch PDF noch präzisere Medienauswahl treffen

8. Weniger Kosten für "falsch" bestellte Literatur (Fernleihe)

4. Nutzen für Informationszentrum

1. Benutzerkreis wird über hochwertiges online-Angebot erweitert - bessere Profilierung
2. Anzahl der überflüssig gelieferten Titel wird stark reduziert - Kosteneinsparung Transport
3. Keine unnötigen Mehrfachexemplare - Kosteneinsparung Investitionen
4. Vorhandene Resourcen werden gezielter genutzt
5. Inhaltserschließung wird preisgünstig, aber sehr nachhaltig vertieft ohne großen Zeitaufwand
6. Erschlossene Titel werden besser "promoted" - Verlage schätzen diese "Werbung"
7. Digitale Kopie wertvoller Exemplare - Bestandssicherung

5.1 Input-Output-Modell von intelligentCAPTURE



5.2 Hohe Performance durch: Multitasking und Multiuser-Verarbeitung

1 **Shell Manager**- steuert das Zusammenspiel der Shells
1 Domino-Server - steuert den Zugriff auf die gemeinsame Datenbank

Automatische Shells

Splitter	- empfängt TIF- und PDF-Image-Datein, zerlegt Multipage-TIF in einzelne Seiten
Zone	- kann Spalten, Bildflächen etc. für OCR markieren, editierbar
Capture	- verwandelt TIF-Dateien mittels OCR - rechenintensiv
QuickFix	- zeigt OCR-Zweifelsfälle, editierbare Qualitätskontrolle
Bind	- setzt die einzelnen Seiten wieder zusammen
PDF-Export	- Produziert eine PDF-Datei und eine Text-Datei
CAI	- ermittelt maschinelle die wichtigsten Deskriptoren - rechenintensiv

Manuelle Schritte

Scanning	- Einscannen der Inhaltsverzeichnisse, Artikel, Kapitel ...
CAI-Check	- manuelle Qualitätskontrolle
Export	- Export an Bibliothekssystem, Dokumentationssystem oder IC CATE

Beispiel: 4 Workstation: einer scannt im Vordergrund und macht Bind und Export im Hintergrund. Eine zweite Workstation macht OCR im Hintergrund und Zone im Vordergrund, eine dritte Workstation macht QuickFix im Vordergrund und CAI im Hintergrund, ein vierter macht CAI-Checking im Vordergrund und lässt Splitter, Bind, PDF-Export und CAI im Hintergrund laufen. Die erste Workstation ist Shell Manager. Der Domino-Server läuft auf dem Netzserver.

6. Maschinelle Indexierung durch CAI-Engine

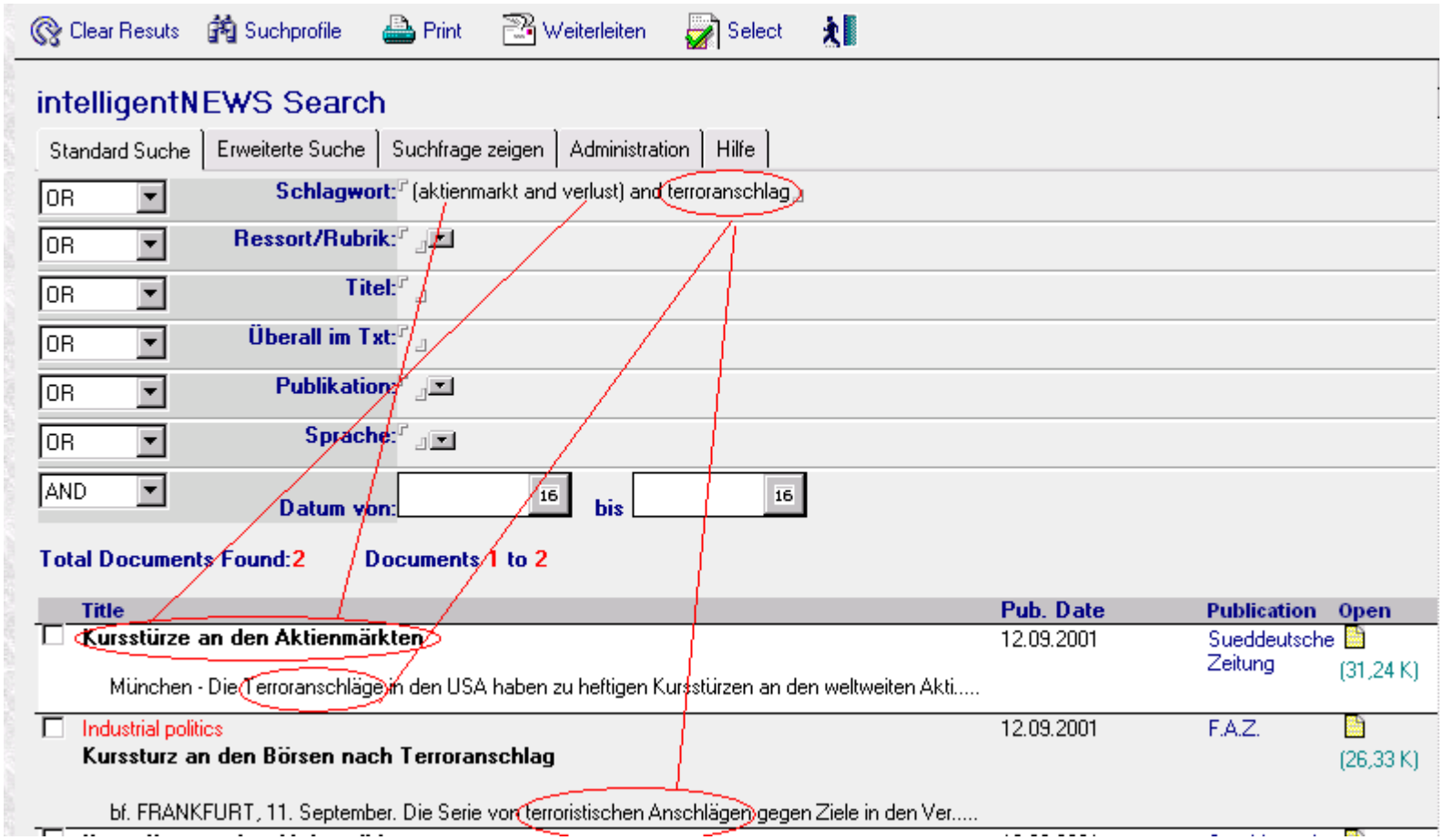
CAI - Computer Aided Indexing

basiert auf AUTINDEX des IAI
Deutsch / English - optional auch Französisch, Spanisch, Italienisch
ca. 200 Mannjahre Entwicklungsleistung

Verfahren:

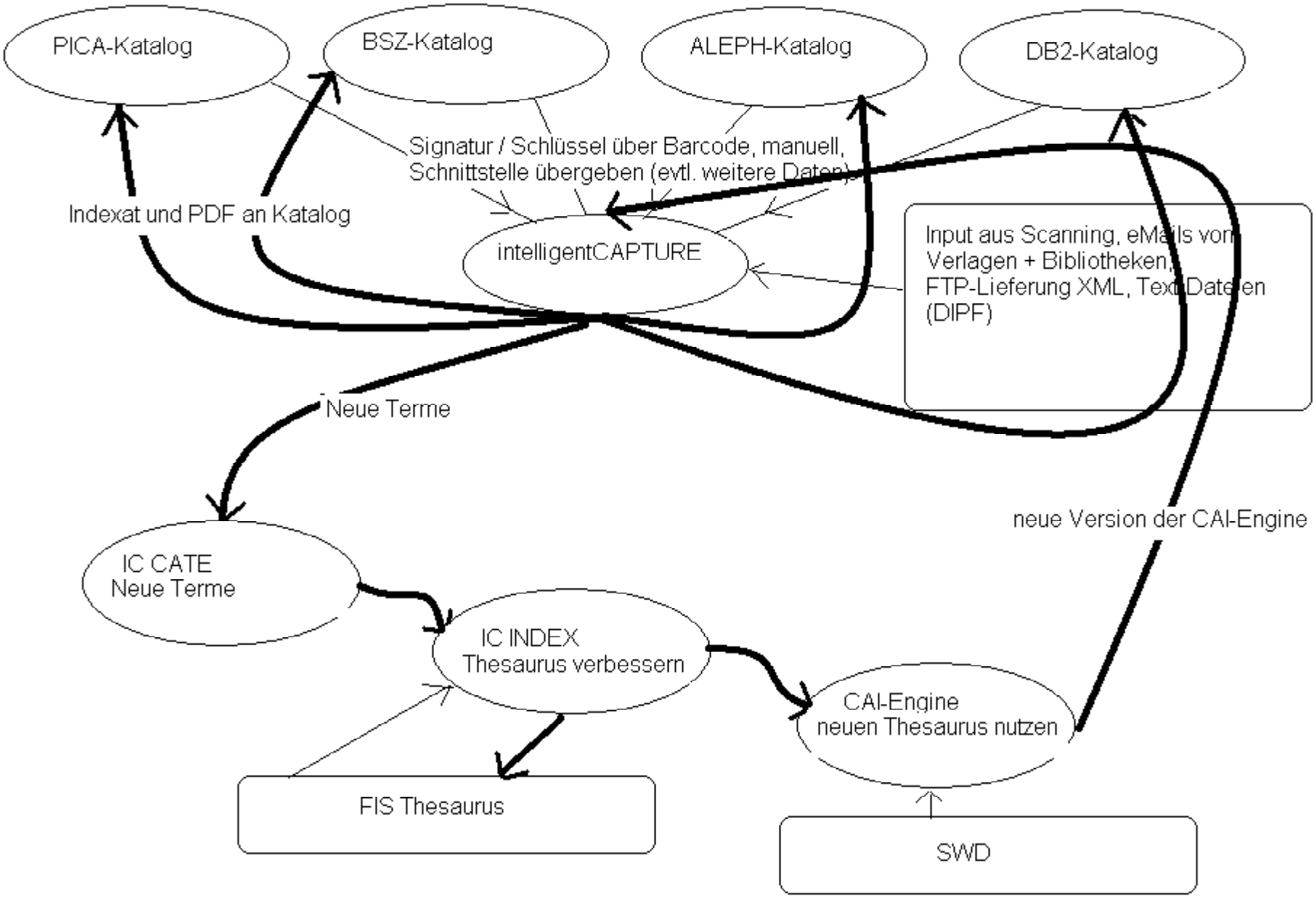
Zerlegt Terme morphologisch -> Grundwort
Erkennt Eigenschaften des Grundwortes
Analysiert Sätze
Analysiert Textstrukturen
Erkennt inhaltlich signifikant wichtige Textworte und Phrasen
Ermittelt zugehörige Deskriptoren aus Thesauri oder Klassifikationen
Ermittelt Personen, Organisationen, Unternehmen
Ermittelt geographische Begriffe
Übersetzt Deskriptoren und wichtige Terme (BINDEX)
Generiert Zusammenfassung

6.1 Beispiel: Sucheffekte durch CAI-Engine



7. Welche Terminologie? Terminologie-Pflege

Terminologiepflege: Projekt beim BMBF


PICA-Katalog
BSZ-Katalog
ALEPH-Katalog
DB2-Katalog
Signatur / Schlüssel über Barcode, manuell,
Schnittstelle übergeben (evtl. weitere Daten)
Indexat und PDF an Katalog
intelligentCAPTURE
Input aus Scanning, eMails von
Verlagen + Bibliotheken,
FTP-Lieferung XML, Text-Dateien
(DIPF)
Neue Terme
neue Version der CAI-Engine
IC CATE
Neue Terme
IC INDEX
Thesaurus verbessern
CAI-Engine
neuen Thesaurus nutzen
FIS Thesaurus
SWD