

NEUROSOFT® 

Neurobook

Inteligentne repozytoria wiedzy

Cezary Dołęga
adres-mailowy@neurosoft.pl

Plan prezentacji

- Definicja repozytorium wiedzy, cechy
- Techniki SI w repozytoriach wiedzy
 - Przetwarzanie obrazów
 - Analiza treści i przetwarzanie języka naturalnego
- Przykłady repozytoriów – projekty Neurosoft
 - Prawo, Normy, Patenty, **Książki**

Definicja

- **Repozytorium** (łac. *Repositorium*) – miejsce **uporządkowanego** przechowywania dokumentów, z których wszystkie są przeznaczone do **udostępniania**, zaprojektowane w taki sposób, aby dostęp do wszystkich jego zasobów był równie **łatwy**.
- **Repozytorium wiedzy** – repozytorium dokumentów z określonej dziedziny, wyposażone w specyficzne dla tej dziedziny mechanizmy ułatwiające dostęp do informacji, lub/i mechanizmy **syntezujące dodatkowe informacje** na bazie treści przechowywanych dokumentów.

Cechy repozytorium wiedzy

- **Kompletność** – chronologiczna, tematyczna, etc...
- **Aktualność** – nowe dane mogą zmieniać stare
- **Powiązania** – odwzorowane relacje między danymi
- **Przeszukiwalność** – FTS, taksonomie
- **Dostępność** – postać cyfrowa, Internet, PDA, ePapier

SI w repozytoriach wiedzy

- **Na etapie opracowywania zawartości**
 - OCR – digitalizacja dokumentów drukowanych
 - Strukturalizacja treści dokumentów
 - Automatyczne wykrywanie powiązań
 - Normalizacja tekstu na potrzeby wyszukiwania
 - Klasyfikacja dokumentów – miejsce w taksonomii
 - Generowanie streszczeń
 - Synteza nowych faktów, ontologie

SI w repozytoriach wiedzy (cd)

- **Na etapie eksploatacji**

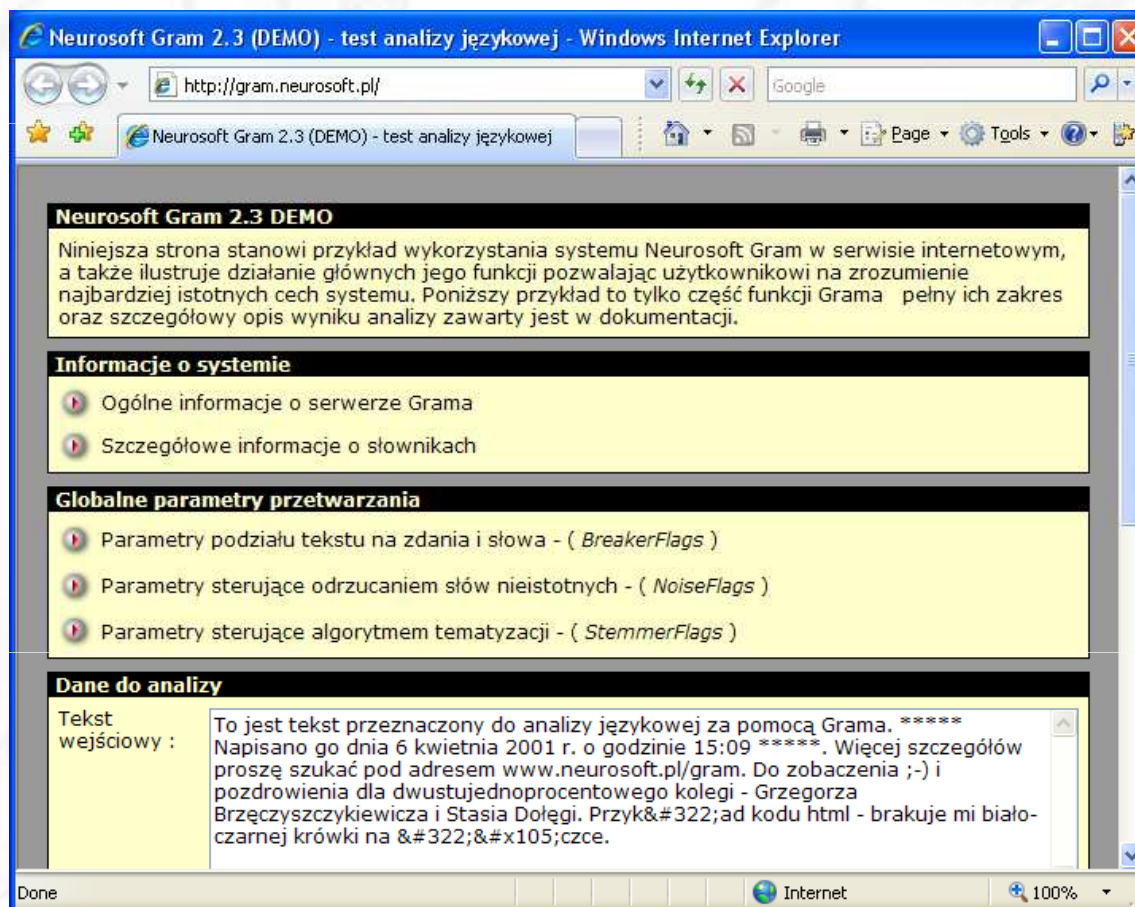
- Analiza zapytań (np. normalizacja)
- Korekta zapytań (ang. „*Did you mean ... ?*”)
- Uszczegóławianie zapytań (ang. „*Suggest ...*”)
- Wyszukiwanie semantyczne
- Systemy dialogowe, konwersacja

Własny OCR – dlaczego ?

- Zła jakość oryginału
- Wysokie wymagania



Normalizacja tekstu – Gram

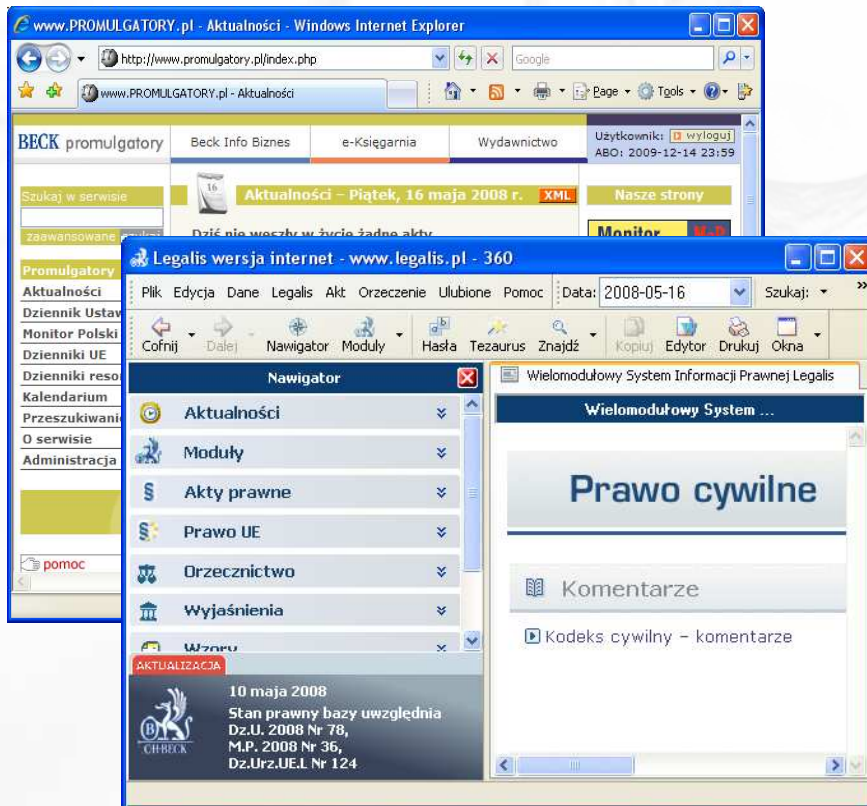


prezentacja on-line

Projekty Neurosoftu

- **Neurolex** – całe prawodawstwo polskie i UE
realizacja od 2001 r. do dzisiaj (C.H.Beck Sp. z o.o.)
- **Neuronorm** – Polskie Normy
realizacja od 2004 r. do dzisiaj (ArsBoni Sp. z o.o.)
- **Neuropat** – patenty i wzory użytkowe
realizacja 2008 r. (Urząd Patentowy RP)
- **Neurolib** – czytelnia internetowa
realizacja 2008 r.

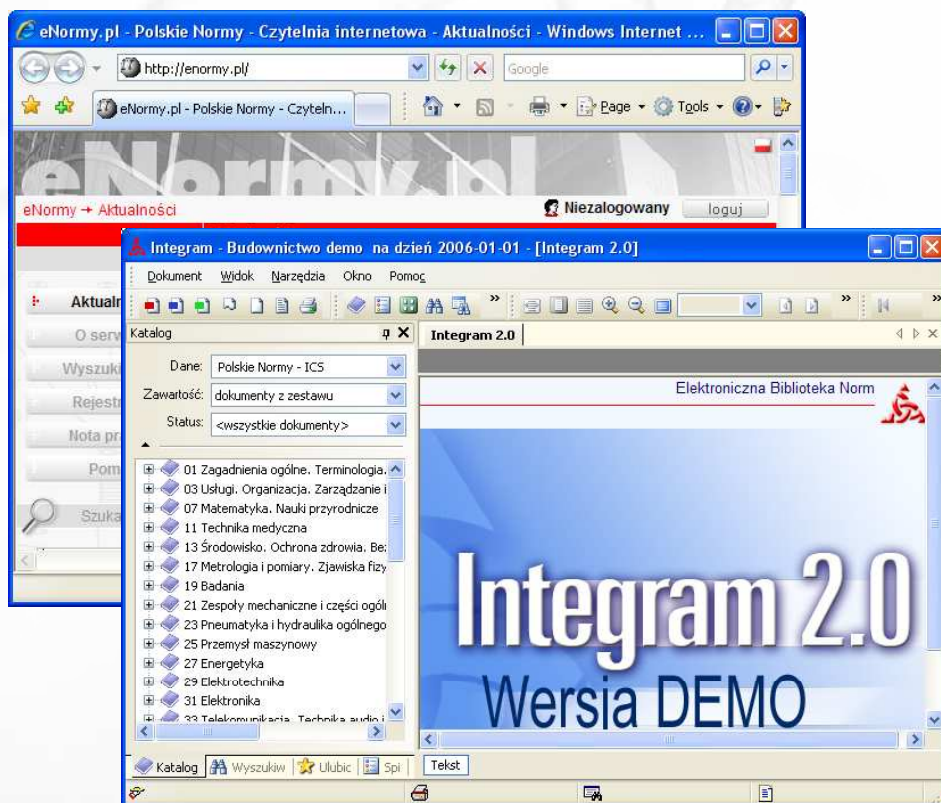
Neurolex



Legalis – system informacji prawnej

- „Papier -> XML + obraz”
- Całe prawo PL i UE (bez lokaln.)
- 928 tys. stron oryginałów
- 282 tys. dokumentów
- 57,5 tys. XML-i z treścią
- Pełna struktura aktu w XML
- Odesłania i cytowania
- Przetwarzanie – codziennie kilkaset stron

Neuronorm

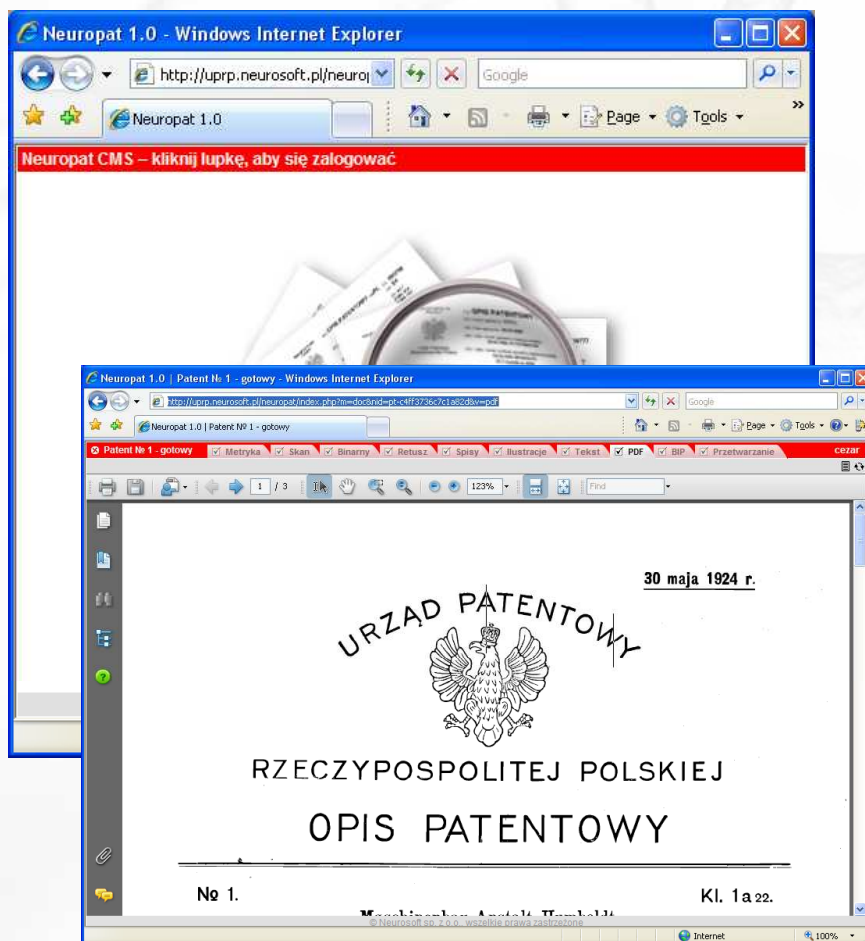


- Wszystkie Polskie Normy
- Źródło = papier + PDF + HTML
- Dokument = obraz + metryka
- Taksonomia ICS
- 583 tys. stron on-line
- ~33 tys. dokumentów on-line
- ~70 tys. metryk dokumentów
- DRM do ochrony kopii

eNormy.pl – Polskie Normy On-line

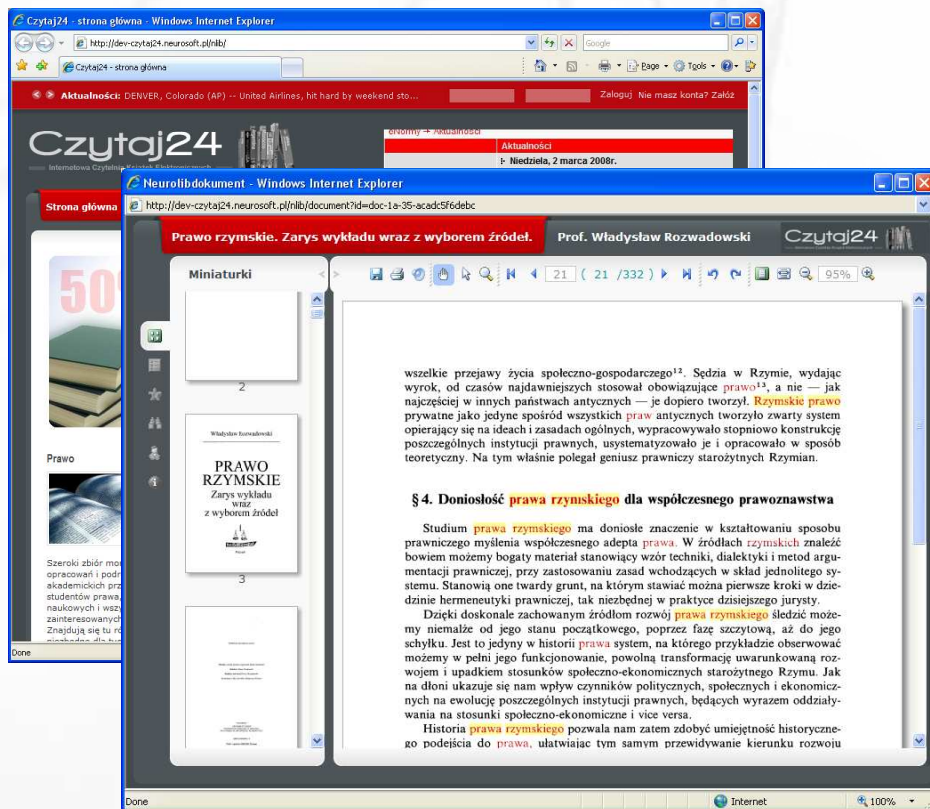
Integram – Polskie Normy na CD

Neuropat



- Polskie patenty i wzory użytkowe
- Lata 1924-1991
- ~780 tys. stron
- 154 tys. patentów, 5 tys. wzorów
- Taksonomia IPC
- Źródło = papier
- Wynik = PDF + txt + metryka
- Wykryte spisy treści

Neurolib



Czytaj24.pl – czytelnia internetowa

- Czytelnia książek on-line
- Źródło = papier lub PDF
- Dokument = obraz + tekst
- Wyszukiwanie pełnotekstowe z fleksyjnym podświetlaniem wyników
- Technologia Java (applet)
- Strukturalizacja wg spisu treści
- Synteza mowy
- Dostęp = PC, telefon, ePapier

Dziękuję za uwagę