dr Mirosław Dąbrowski
mgr inż. Lucja Rzęcka

M. Dąbrowski.

Kierunki rozwoju systemów informacyjnych

/ referat pomocniczy dla Podsekcji Informatyki
Sekcji Informatyki, Automatyki i Pomiarów
II Kongresu Nauki Polskiej /

Październik, 1972
Warszawa

Spis treści:

1. Uwagi wstępne

Celem niniejszego opracowania jest analiza stanu prac w kraju i zagranicą nad zastosowaniem maszyn cyfrowych do procesów informacyjnych oraz podanie głównych kierunków rozwoju zautomatyzowanych systemów informacji, obserwowanych już w chwili obecnej za granicą. Za podstawę analizy przyjęto istniejące systemy informacyjne, albowiem znane nam projekty nowo powstających systemów nie odbiegają w swoich rozwiązaniach od istniejących już, a poza tym w trakcie realizacji systemy często projekty te ulegają zmianom. Z tych też względów nie będziemy oceniać opracowywanego obecnie Krajowego Systemu Informacyjnego, natomiast wydaje się nam, że zasygnalizowane tutaj tendencje rozwoju zautomatyzowanych systemów informacyjnych winny być brane pod uwagę przy opracowywaniu Krajowego Systemu Informacyjnego.

Nie będziemy tu również omawiać wszystkich istniejących systemów informacyjnych za granicą, omówimy jedynie te systemy, które w latach ubiegłych odgrywały ważną rolę w rozwoju automatycznych systemów informacyjnych, bądź w chwili obecnej należą do najnowocześniejszych systemów na świecie.

Szczególna uwaga będzie w niniejszym opracowaniu zwrócona na zagadnienia związane z gromadzeniem i wyszukiwaniem informacji, gdzie informacja ma postać dokumentu

Wyszukiwanie informacji jest dosyć nową dziedziną badań zajmującą się zastosowaniem maszyn cyfrowych dla potrzeb służby informacyjnej. W kontekście wyszukiwania informacji jest to proces, którego zakres może sięgać do przekazywania informacji między szukającym jej a kartoteką kart do dialogu człowieka z maszyną. Dlatego też ważnym jest dobór języka informacyjnego oraz odpo-

wiednia organizacja zbioru informacji. Ważnym zagadnieniem w systemach informacyjnych jest również problem opracowywania dokumentu (automatyczne indeksowanie) jak i samo przekazywanie dokumentu oryginalnego (kartoteki mikrofilmów itp.).

Celem automatyzacji systemów informacyjnych jest:

- przyśpieszenie opracowywania dużych zbiorów dokumentów,
- skrócenie czasu wyszukiwania informacji,
- zmniejszenie kosztów gromadzenia i wyszukiwania informacji, przy czym efekty ekonomiczne opracowania informacji uzyskuje się dopiero w okresie gdy automatyczny system informacyjny funkcjonuje już sprawnie,
- osiągnięcie jakościowo lepszych wyników opracowań informacyjnych oraz różnego typu indeksów,
- zapewnienie obiektywności opracowywania i wyszukiwania informacji.

Głównym celem systemu wyszukiwawczego jest odszukanie i maszynowe sporządzenie tematycznych zestawień w/g żądanych profili (wyszukiwanie selektywne) czy też kwerend (wyszukiwanie retrospektywne) na podstawie zbiorów dokumentacyjnych zgromadzonych i zapamiętanych na maszynowych nośnikach informacji. "Kluczem" wyszukiwania w nowoczesnych systemach informacyjnych jest kontrolowany język deskryptorowy zawarty w tezaurusie. Przy czym pod pojęciem "tezaurus" rozumiemy wyselekcjonowany i uporządkowany zbiór słów przyjętych do indeksowania tzw. deskryptorów lub słów kluczowych. W tezaurusie istnieją relacje synonimiczności, a czasem nawet relacje hierarchii i kojarzeń, które ułatwiają w sposób jednoznaczny indeksowanie dokumentów i pytań.

Bardzo często w systemie wyszukiwawczym wykonuje się dodatkowe operacje na zbiorach dokumentów w emc (programy badań statystycznych) oraz programy wydawnicze), które umożliwiają okresowe wyda-

wanie "Informatorów" i różnego rodzaju indeksów, tezaurusa itd. Systemy takie nazwane są systemami wyszukiwawczo-wydawniczymi (przykładem jest system INIS). Istnieją także systemy pełniące funkcje tylko wydawniczą, których głównym zadaniem jest operowanie formatem danych, modyfikacje i przygotowanie (wydawnictwo) katalogów.(Przykładem jest system MARC - MAchine Readable-Catalog)

2. Analiza istniejących systemów informacyjnych

2.1. Analiza krajowych systemów informacyjnych

Pierwsze prace nad systemami informacyjnymi w Polsce rozpoczęto w 1966 roku. Miały one charakter eksperymentalny i kończyły się często na próbach, były powtórzeniami prac już wcześniej opisanych lub stanowiły obiekt ulepszeń i rozwiązań własnych. Przykładem mogą być prace Politechniki Wrocławskiej (system Automatycznej Selekcji Patentów i zamiany z klasyfikacji amerykańskiej na klasyfikację niemiecką) oraz prace Instytutu Maszyn Matematycznych (system INBI[1] i IBIS). Podstawowym celem systemu INBI miało być usprawnienie istniejącego już wydawnictwa informacyjnego pod tytułem "Biuletyn Informacyjny IMM".

Podobnemu celowi służą później opracowywane systemy: IGA - wydawnictwo"Informatora o zakończonych pracach naukowo-badawczych", system KWOC - wydawnictwo "Przegląd Piśmiennictwa i Zagadnień Informacji" oraz opracowany ostatnio system ARKA - dla automatycznego redagowania "Katalogu Czasopism Zagranicznych" w Bibliotece Narodowej.

Niektóre elementy opracowane w systemach IGA i KWOC pozwoliły na stworzenie i zbadanie w naszym warunkach dwóch podstawowych indeksów permutacyjnych typu KWIC i KWOC.

---

[1] Omawiane w tym punkcie skróty systemów są wyjaśnione w Załączniku 1. Tam też są podane dokładniejsze dane o tych systemach.

W systemie IGA został opracowany indeks KWIC w dwóch wersjach:

- obcinający tekst do pojedyńczego wiersza oraz
- zawijający wiersz.

Ponadto, system IGA daje możliwość sporządzania indeksów zbiorczych oraz wyszukiwania informacji na żądanie wg takich cech jak:

- symbol UKD, nazwisko autora lub haseł.

Natomiast system KWOC jest pierwszym systemem w Polsce, w którym zastosowano metodę automatycznego kojarzenia (porównywania) przez maszynę cyfrową słów kluczowych (opracowanych uprzednio ręcznie) z występującymi w tytułach mutacjami gramatycznymi tych wyrazów. Metoda ta stanowi ułatwienie oraz przyśpieszenie prac związanych z przyporządkowaniem poszczególnym słowom kluczowym odpowiadających im wyrazów w tekście, które to czynności są szczególnie pracochłonne i żmudne w wykonaniu ręcznym.

Drugim systemem, w którym podjęto badania uwzględniające specyfikę języka polskiego pod względem jego właściwości do celów automatycznego przetwarzania tekstów był system MARYSIA opracowany na Uniwersytecie Warszawskim.

Badania te mają szczególne znaczenie, bowiem liczne prace i wyniki prób przeprowadzone na świecie dotyczyły języka angielskiego. Fleksja języka angielskiego występuje w postaci szczątkowej i jest znacznie prostsza aniżeli fleksja języków romańskich czy germańskich, natomiast jest bardzo złożona w językach słowiańskich. Dlatego też analiza fleksyjna tekstu polskiego ma dla nas szczególne znaczenie, kontynuacja tych badań wydaje się być koniecznością.

W zakresie języków informacyjnych dotychczas wykonano w Polsce kilkanaście tezaurusów, przy czym szczególny nacisk położono na uwzględnienie relacji hierarchicznych między deskryptorami.

Tezaurusy te były wykonane ręcznie i niezależnie od systemów informacyjnych. Do chwili obecnej żaden system eksploatowany w Polsce nie ma tezaurusa w pamięci maszyny cyfrowej.

Tak postawiony problem budowy tezaurusów, a ogólnie języków informacyjnych może znacznie zahamować rozwój systemów informacyjnych w kraju.

Systemami wyłącznie wyszukiwawczymi są systemy IBIS i SAWI-2P. Wyszukiwanie zaproponowane w systemie IBIS przewiduje każdorazowe przeszukiwanie wszystkich dokumentów (zarówno części klasyfikacyjnej jak i opisowej) jest to bardzo czasochłonne (8 min. w zbiorze 1000 dokumentów), a ponadto powoduje duże szumy informacyjne. System ten poza kilkoma pokazami nie był nigdy praktycznie wykorzystany.

Bardzo ciekawym systemem wyszukiwawczym jest stosunkowo najwcześniej opracowany system APIS. Po raz pierwszy w Polsce zastosowano odmienną technikę wyszukiwania, mianowicie poprzez kartoteki inwersyjne. Jednak wadą systemu APIS jest używanie symboli numerycznych Branżowej Klasyfikacji i nazw jednostek pływających do indeksowania dokumentów i tworzenia zbiorów (jest to wynikiem ograniczenia spowodowanego konfiguracją emc Eliott-803 o bardzo małej pamięci). Technika wyszukiwania oparta na zastosowaniu kartotek inwersyjnych (różnego rodzaju) jest w chwili obecnej stosowana we wszystkich nowoczesnych systemach na świecie. Ulepszoną kartotekę inwersyjną używa także system SAGO-CZAS, który obok systemu ASIA jest najbardziej rozwiniętym systemem wyszukiwawczo-wydawniczym.

System SAGO-CZAS oparty jest na języku deskryptorowym kontrolowanym pszczmocą tezaurusa oraz posiada pewne elementy gramatyki. Natomiast system ASIA posiada bardzo starannie opracowane wydawnictwa. Do indeksowania dokumentów w systemie ASIA

używany jest język deskryptorowy z niekontrolowanym słownictwem, brak opracowanego do chwili obecnej tezaurusa praktycznie uniemożliwia wyszukiwanie dokumentów. Zastosowana technika wyszukiwania polega na przeszukaniu kolejno wszystkich dokumentów w zbiorze (tj. bardzo czasochłonne), ale pozwala na zapytanie dotyczące każdego pola notki, a więc, można pytać nie tylko o treść, ale również o autora, język, kraj i wiele innych niewykorzystywanych elementów opisu bibliograficznego.

Pewne elementy gramatyki opisu deskryptorowego posiadają systemy ASIA i SAGO-CZAS. W systemach tych do formułowania pytań używane są operatory logiczne "i" oraz "lub", a w systemie ASIA - operator logiczny "negacja". Natomiast brak jest systemu umożliwiającego używanie operatorów porównania: " ", " = ", " " przy wyszukiwaniu wartości liczbowych cech.

Prawidłowy rozwój systemów informacyjnych w kraju wymaga prowadzenia badań podstawowych w tym zakresie. Badania takie w Polsce są prowadzone od kilku lat. Dotyczą one matematycznych podstaw systemów wyszukiwania informacji, a w szczególności modeli matematycznych systemów wyszukiwania informacji, struktur danych, organizacji dużych zbiorów informacji. Kontynuacja tych badań jest konieczna jak również koniecznym wydaje się prowadzenie badań nad językami informacyjnymi jak i zastosowaniem do nich lingwistyki matematycznej.

Na zakończenie oceny sytuacji krajowej należy stwierdzić brak wymiany informacji w Polsce na temat prac prowadzonych z omawianego tu zakresu. Dotychczas nie została zorganizowana żadna konferencja lub szkoła letnia dotycząca systemów informacyjnych. W wielu przypadkach mamy do czynienia z pracami "od początku".

2.2. Analiza zagranicznych systemów informacyjnych

Początki badań nad systemami wyszukiwania informacji za granicą dotyczą połowy lat 50-tych. Pierwsze systemy wyszukiwania informacji były tworzone dla pewnych dziedzin zastosowań i tak w 1964 roku system MEDLARS (Medical Literature Analysis and Retrieval System) był zastosowany w medycynie, system MARLIS (Multi-Aspect Relevance Linage Information System) zastosowano w chemii, system MASIS (Management and Scientific Information System) zastosowano do zarządzania pracami naukowo-badawczymi, a system ALTAIR (Automatic Logical Translation and Information Retrieval) zastosowano w astronomii. Kolejne wersje tych systemów były udoskonalane, rozszerzane i przystosowywane do coraz większych możliwości maszyn cyfrowych.

Oprócz tych systemów powstawały systemy wyszukiwania informacji ogólnych zastosowań. Były to systemy firmowe takie jak np. system NIC firmy ICL, system FIND na maszynach cyfrowych serii 1900, system MISTRAL firmy IRIS, system IRMS oraz STAIRS firmy IBM, system GOLEM firmy SIEMENS. Ten ostatni system (GOLEM) jest największym i najbardziej nowoczesnym systemem wyszukiwania i przekazywania informacji. Prace wstępne nad tym systemem rozpoczęto w 1965 r. a od 1969 r. jest on stosowany na uniwersytetach i w ośrodkach informacji i dokumentacji w wielu krajach. W 1972 r. system ten został zastosowany do obsługi igrzysk olimpijskich w Monachium. System GOLEM pracujący na maszynie Siemensa 4004/45 pozwala na przekazywanie informacji o liczebności 4 mld. znaków i podzielonej na co najwyżej 150 zakresów tematycznych. System ten wykorzystuje pamięci masowe o dostępie bezpośrednim tj. dyski (do przechowywania tezaurusa i kartotek inwersyjnych) oraz karty magnetyczne, na których umieszczone są opisy dokumentów, Nie wy-

korzystuje natomiast taśm magnetycznych. System GOLEM eksploatowany jest w oparciu o wielodostępny układ teletransmisji umożliwiający jednoczesne użytkowanie 36 stacji display'owych. System ten może zawierać do 300000 deskryptorów. Proces wyszukiwania na pytanie 10-cio deskryptorowe w zbiorze 1 miliona dokumentów trwa maksymum 30 sek.

W krajach zachodnich zarówno prace o charakterze teoretycznym jak i prace dotyczące realizacji konkretnych systemów wyszukiwania informacji są bardzo zaawansowane. Prawie wszystkie prace teoretyczne dotyczące modeli systemów wyszukiwania informacji, a w szczególności nowych rozwiązań tych systemów zostały sprawdzone na maszynach cyfrowych. Przykładem tych prac jest jeden z pierwszych, klasyczny system analizy tekstów i wyszukiwania informacji - system SMART (Salton's Magical Automatic Retrieval of Texts). Został on opracowany przez Saltona w 1964 r. w ośrodku obliczeniowym Uniwersytetu Harward i przeznaczony do przebadania różnych algorytmów analizy tekstu i wyszukiwania informacji na żądanie. Algorytm automatycznej analizy tekstu wykorzystuje wagi oraz współczynniki korelacji poszczególnych słów tekstu. Próby przeprowadzone na automatycznie utworzonym tezaurusie i na tezaurusie opracowanym ręcznie wypadły pomyślnie. Dokumenty w pamięci maszyny cyfrowej były rozmieszczone w tzw. wiązki, co znacznie zmniejszyło czas wyszukiwania, albowiem badanie dokumentów będących odpowiedzią na zadane pytanie było przeprowadzane tylko dla pewnych "wiązek" gdzie w całym zbiorze dokumentów.

W ramach systemu SMART powstał system SOOCER (SMART's own Coordance Constructor, Extremely Rapid), który umożliwił szybką konkordancję tzn. tworzenie słownika do tekstu z wykazaniem dla każdego słowa jego położenia w tekście.

Innym przykładem prac teoretycznych, które zostały sprawdzone na maszynie cyfrowej był model matematyczny systemu wyszukiwania informacji zaproponowany przez S. P. Gosha i C. T. Abrahama, a oparty na geometrii skończonej nad polem Galoisa. Został on sprawdzony na maszynie IBM-360.

Dotychczas ukazały się "setki" publikacji dotyczących różnych aspektów systemów wyszukiwania informacji, od badań podstawowych, modeli systemów, formalizacji procesu wyszukiwania, poprzez analizę istniejących systemów do omówienia konkretnych realizacji systemów włącznie.

Dokładniejsze dane dotyczące systemów zagranicznych są zawarte w załączniku Nr 2. Dane dotyczące pewnych cech wspólnych nowoczesnych systemów wyszukiwania informacji będą podane w następnym punkcie niniejszego opracowania. Oprócz badań nad wykorzystaniem języków deskryptorowych (tezaurusów) do systemów wyszukiwania dokumentów prowadzone są również prace nad wykorzystaniem Uniwersalnej Klasyfikacji Dziesiętnej (UKD) w systemach wyszukiwawczych. Zasięg tych prac jest znacznie mniejszy aniżeli prac nad językami deskryptorowymi. Tym nie mniej, szczególnie w Stanach Zjednoczonych przeprowadzane są eksperymenty (np. przez Fremana i Ahertona) polegające na przekształceniu zapisu klasyfikacji, tak aby był on w sposób efektywny wykorzystywany w systemach wyszukiwania informacji. Do takich systemów należą np. AUDACIONS i LACS. Przykładowe systemy wyszukiwania informacji oparte na języku UKD są zawarte w załączniku 3. Na zakończenie tego punktu wypada wspomnieć o systemach typu konwersacyjnego, które również mogą pełnić funkcje wyszukiwawcze. Pozwalają one na konwersację człowieka z maszyną cyfrową w języku naturalnym. Do takich systemów należy zaliczyć system REL (Rapidly Extensible System Language), wykonany w California Institute of Technology, oraz system konwersacyjny wykonany w Massachusetts

Institute of Technology. Wydaje się, że ten ostatni wykorzystujący w mniejszym stopniu analizę syntaktyczną zdań (pytań), oparty natomiast na "sztucznej inteligencji" jest rewelacją w omawianej tu dziedzinie.

3. Kierunki rozwoju systemów informacyjnych

Analiza nowoczesnych systemów wyszukiwania informacji pozwala na wyciągnięcie następujących wniosków charakteryzujących nowoczesne systemy:

a) Systemy zapewniają wyszukiwanie retrospektywne jak i profilowe.

b) Formułowanie zapytań za pomocą trzech operatorów logicznych ("i", "lub", "negacja"), którymi powiązane są słowa kluczowe (systemy: IRMS, STAIRS, GOLEM itp.), a także za pomocą operatorów porównania typu "=", "≠", " ", " " wykorzystywanych do danych przyjmujących wartości liczbowe (np. trzecia wersja systemu MISTRAL).

c) Systemy zapewniają dwuetapową procedurę wyszukiwawczą. Na pierwszym poziomie użytkownik otrzymuje ilość odpowiedzi na zadane pytanie i w zależności od tego może pytanie bardziej uściślić dołączając dodatkowe deskryptory lub rozszerzyć. Dopiero na drugim poziomie użytkownik otrzymuje dokumenty.

d) Możliwość równoczesnego korzystania z kartotek przez wielu użytkowników. Takie systemy jak STAIRS, MISTRAL, GOLEM posiadają szereg urządzeń końcowych, monitorów, zapewniających wielodostępność. Praca w real-time.

e) Wprowadzenie do systemu tzw. "kartoteki tajności" zabezpieczającej dostęp do pewnych zbiorów informacji tylko dla upoważnionych do tego użytkowników. Przeważnie problem ten jest rozwiązany

w taki sposób, że każdy z użytkowników otrzymuje hasło, które pozwala na przekazanie mu jedynie informacji o określonym stopniu tajności.

f) Elastyczność systemu - system może być adoptowany do różnych ośrodków informacji.

g) Możliwość aktualizacji informacji jak i tezaurusa (system IRMS).

h) Systemy zapewniają automatyczną korekcję błędnych słów kluczowych (np. system INIS).

i) Systemy wyszukiwania wykonują dodatkowo następujące czynności:

- badania statystyczne (częstotliwość pojawiania się słów kluczowych,
- okresowa edycja tezaurusa lub indeksu,
- sporządzanie katalogów,
- opracowanie tematycznych biuletynów sygnalnych.

j) Systemy umożliwiają automatyczne tworzenie słownika w oparciu o również automatyczną analizę dokumentów (np. system NEW YORK TIMES).

k) Systemy są realizowane na najnowocześniejszych maszynach cyfrowych.

l) Systemy operują na dużych zbiorach deskryptorów i na dużych zbiorach informacji.

Podane wyżej cechy cechy charakteryzujące nowoczesne systemy wyszukiwania informacji zabezpieczają pełną automatyzację procesu gromadzenia i wyszukiwania informacji, elastyczność i uniwersalność systemu.

W aspekcie wyżej wspomnianych cech nowoczesnych systemów wyszukiwania informacji, nasze systemy krajowe pozostają daleko w tyle. Realizacja krajowych systemów informacyjnych o standarcie

światowym nie może być dokonywana tylko drogą zakupu i adaptacji całych zachodnich systemów. Nie pozwala na to leksyka języka polskiego, która jest bardzo złożona w porównaniu z leksyką języka angielskiego. Wydaje się koniecznym szersze prowadzenie badań nad formalizacją języka polskiego, nad jego leksyką, analizą syntaktyczną frez etc.

Należy dodać, że pierwsza próba uruchomienia systemu zagranicznego w Polsce została podjęta w COKB w Gdańsku w 1967 r. Był to system firmowy ICL FIND (wersja I) oraz system NIC. Jednak ograniczenia systemu i skomplikowany sposób operowania spowodowały rezygnację z dalszych prób nad tym systemem. Doświadczenia te i dokładna analiza tego systemu zostały wykorzystane do opracowania techniki wyszukiwania w systemie ASIW, a następnie w systemie SAGO-CZAS.

W chwili obecnej prowadzone są prace nad wykorzystaniem systemu IRMS dla potrzeb Polskiego Radia i Telewizji.

Drugą sprawą mogącą znacznie opóźnić rozwój krajowych systemów wyszukiwania informacji jest niewłaściwie ukierunkowany rozwój języków deskryptorowych. Do chwili obecnej opracowano w kraju kilkanaście tezaurusów. Opracowywane są one w sposób ręczny. Należy zdać sobie sprawę z roli tezaurusa w systemie wyszukiwania informacji. Tezaurus w systemie takim jak np. MISTRAL lub GOLEM jest po prostu słownikiem słów kluczowych, wyselekcjonowanych, uszeregowanych i połączonych odpowiednimi relacjami (nie koniecznie hierarchicznymi). Powstaje on równocześnie z systemem wyszukiwania i jest przez ten system tworzony i ciągle uzupełniany. Tezaurus nie jest systematyką słownictwa i jeszcze jedną jego "sztywną" klasyfikacją, jest on jedynie językiem dla efektywnego maszynowego przetwarzania i wyszukiwania informacji. Tezaurus jest tylko narzędziem systemu informacyjnego, a nie jest celem samym w sobie.

Załącznik Nr 1

Charakterystyka krajowych systemów informacyjnych

I. Systemy eksploatowane

Lp.	Nazwa systemu	Wykonawca	Rok opracowania	Typ EMC	Moc zbioru deskryptorów	Moc zbioru informacji
1.	Automatyczne Poszukiwanie Informacji Syntetycznej (APIS)[1]	COKB Przemysłu Okrętowego Gdańsk	1968	Eliott		ok.250
2.	Automatyczny System Informacji o wyjazdach służbowych za granicę (ASIW)[2]	CPITE	1969	ICL-1904		ok. 2000
3.	Automatyczna Selekcja Informacji Adresowej (ASIA)[3]	UNITECH	1970	ICL-1904		ok. 10 000
4.	System Automatycznego Gromadzenia i Wyszukiwania (SAGO-CZAS)[4]	CPITE	1970	ICL-1904		ok. 7000
5.	Automatyczna Selekcja Patentów[5]	Politechnika	1967			

1 System wyszukiwawczy. Obejmuje bibliografię artykułów publikowanych w okętowej i technicznej prasie światowej. System dostosowany do bardzo specjalistycznej informacji. Ze względu na stosowany sposób indekscwania dokumentów (symbole numeryczne) jak również formę wydawniczą, system jest użyteczny dla wąskiego kręgu odbiorców. Nie przewiduje się ulepszenia lub rozszerzenia systemu.

2 System wyszukiwawczo-wydawniczy. Obejmuje sprawozdania z wyjazdów zagranicznych dwóch resortów. Pytania kierowane do systemu mogą dotyczyć zarówno danych formalnych o wyjeżdżającym jak nazwisko, kraj, firma itd. jak i treści sprawozdań. System spełnia także funkcje kontroli terminowości składania sprawozdań, wydaje monity. Ponadto kwartalnie wydawany jest Informator o wyjazdach zagranicę z indeksami.

3 Celem systemu jest przyśpieszenie prac redakcyjnych "Informacji bieżądej" oraz gromadzenie i wyszukiwanie informacji z zakresu elektroniki i teletechniki. Jest to system wydawniczo wyszukiwawczy. Dokumenty opisuje się deskryptorami z zastosowaniem uproszczonej gramatyki (deskryptory główne, punkty widzenia: ogólny i uściślający, deskryptory zwykłe). Podczas eksploatacji systemu prowadzi się ewidencję deskryptorów. Główny nacisk położono na redakcję "Informacji bieżądej". Technika wyszukiwania polega na przeszukiwaniu całego zbioru, co jest bardzo czasochłonne.

4 System wydawniczo-wyszukiwawczy. Zarówno dokumenty jak i pytania opisuje się deskryptorami zebranymi w branżowych tezaurusach, co zwiększa dokładność wyszukiwania (usunięcie synonimii i ujednolicenie terminologii). Do opisu dokumentów wprowa-

dzono uproszczoną gramatykę (deskryptory tematyczne i uściślające, wyróżnienie deskryptorów głównych). Technika wyszukiwania stosująca ulepszoną kartotekę inwersyjną jest korzystna przy dużych zbiorach i dużym zapotrzebowaniu na informacje. Celem systemu jest centralizacja gromadzenia, wyszukiwania i wydawania informacji w resortach przemysłu ciężkiego i maszynowego. Jest to system doświadczalny, eksploatowany obecnie na danych z dwóch ośrodków branżowych: CBKUB i CBKM (transport bliski i urządzenia budowlane).

[5] System ten dokonuje automatycznej selekcji patentów oraz dokonuje zamiany z klasyfikacji amerykańskiej na klasyfikację niemiecką.

II. Systemy opracowane do automatyzacji wydawnictwa jednorazowego

Lp.	Nazwa systemu	Wykonawca	Rok opracowania	Typ EMC	Moc zbioru deskryptorów	Moc zbioru informacji
1.	Informacja Grupowania Automatycznego (IGA)[1]	CIINTE i ZOWAR	1967	IBM 1440		ok. 1500
2.	Key-word-out-of-context (KWOC)[2]	CIINTE	1969	ZAM-41		ok. 5000
3.	Automatyczne Redagowanie Katalogów (ARKA)[3]	Biblioteka Narodowa przy współpracy Instytutu Maszyn Mat.	1971	ZAM-41		ok. 1600 docelowo - 16000)

[1] System wydawniczy. Służy do automatycznego sporządzania egzemplarza wydawniczego "Informator o zakończonych pracach naukowych i naukowo-badawczych". Razem z informatorem wydaje kilka indeksów: autorski, przedmiotowy, instytucji, czasopism oraz indeksu KWIC według tytułu oraz spis treści według UKD. Wyszukiwanie przeprowadza się jedynie w celach pokazowych na małym zbiorze. Obecnie zaniechano umieszczania w karcie dokumentacyjnej haseł przedmiotowych, co eliminuje wyszukiwanie.

[2] System dokonuje automatycznego sporządzania indeksu permutacyjnego typu KWOC w oparciu o tytuły polskie publikacji w powiązaniu z wykazami bibliograficznymi, indeksem autorskim i indeksem czasopism. Przedmiotem systemu było 5 roczników "Przeglądu Piśmiennictwa Informacji" wydawanego przez CIINTE w latach 1962-1967. System wykorzystuje następujące zbiory informacji:

- EWIDENCJA - zbiór dokumentów uporządkowanych według działów, których kolejność jest narzucona symboliką UKD. Wewnątrz tych działów dokumenty są rozmieszczone według rosnącej numeracji roczników.

- WYKAZ PAR/SŁÓW KLUCZOWYCH I SŁÓW TEKSTOWYCH - zbiór ten jest opracowany ręcznie a następnie wprowadzony do maszyny cyfrowej.

- INDEKS KWOC - zbiór ten zawiera uporządkowane alfabetycznie słowa kluczowe z numerami identyfikującymi dany dokument w zbiorze EWIDENCJA i w zeszycie PPZI oraz tytuł polski.

- INDEKS AUTORSKI - zbiór dokumentów zawiera nazwisko autora z numerami dokumentów.

- INDEKS CZASOPISM - zbiór zawiera skróty czasopism z numerami dokumentów.

- SŁOWNIK - zawiera wszystkie słowa z tytułów polskich publikacji.

Wszystkie wyżej wymienione zbiory umieszczone są na taśmie magnetycznej.

System KWOC stanowi pierwszą próbę opracowania indeksu permutacyjnego typu KWOC przy użyciu maszyn cyfrowych w Polsce.

3 Celem systemu jest przyśpieszenie cyklu wydawnictwa i poprawa jakości katalogu czasopism zagranicznych. Dokumenty indeksuje się na istniejących kartach bibliotecznych. Kluczem identyfikującym każdy nowy dokument wprowadzony do maszyny cyfrowej jest pierwsza pozycja poprzedzona identyfikatorami TY po czym następuje tytuł czasopisma. Następne identyfikatory tzn. MW - miejsce wydania czasopisma, IS - instytucja sprawcza, JE - język, KR - kraj wydania, SI - miejsce przechowywania czasopisma, OD - odsyłach.

III. Systemy niesksploatowane

Lp.	Nazwa systemu	Wykonawca	Rok opracowania	Typ EMC	Moc zbioru deskryptorów	Moc zbioru informacji
1.	Informacje Bibliograficzne (INBI A)[1]	Instytut Maszyn Matematycznych	1967	ZAM-21		ok.250
2.	Informacje Bibliograficzne (INBI B)[2]	Instytut Maszyn Matematycznych	1968	ZAM-41		ok.250
3.	Informacji Bibliograficznej Szukanie (IBIS)[3]	Instytut Maszyn Matematycznych	1968	ZAM-41		ok.2500
4.	System Automatycznego Wyszukiwania Informacji Patentowej (SAWI-2P)	CIINTE	1970	MIŃSK-22		ok.150

[1] Jest to jeden z pierwszych systemów. Przeznaczony do prowadzenia aktualnej ewidencji pozycji bibiliograficznych z dziedziny maszyn matematycznych. Wykorzystywany do redagowania miesięcznika "Automatyzacja Przetwarzania Informacji - Bibliografia". System wykonany w celach szkoleniowych, po rocznej eksploatacji wycofany z użycia. Wyszukiwanie polegało na przeglądaniu całego zbioru i porównania pytania z nagłówkami dokumentów.

[2] Jest to system INBI A, opracowany na maszynę ZAM-41.

[3] System wyszukiwawczy zademonstrowany na targach w Lipsku. Nie uwzględnia modyfikacji zbiorów. Technika wyszukiwania zastosowana w systemie jest bardzo czasochłonna i może powodować "szumy" informacji.

[4] System wydawniczo-wyszukiwawczy. Nie był eksploatowany. Był uruchomiony na 50 dokumentach. Służy do gromadzenia takiej charakterystyki patentu, która umożliwiłaby wyszukiwanie poprzez różne klasyfikacje i deskryptory. Zastosowana technika wyszukiwania - czasochłonna, czas wyszukiwania wzrasta liniowo wraz ze wzrostem zbioru patentów.

Załącznik Nr 2

Charakterystyka zagranicznych systemów informacyjnych

I. System INIS - Międzynarodowa Agencja Energii Atomowej.

a) system wyszukiwawczo-wydawniczy informacyjny

b) typ EMC - IBM-360/40. Konfiguracja: 64k, 12TM, 2DM, czytnik taśmy i kart, drukarka wierszowa

c) indeksowanie dokumentów - odbywa się przy pomocy najbardziej szczegółowych właściwych deskryptorów. Natomiast podczas wprowadzania do EMC, wszystkie deskryptory szersze są dopisywane automatycznie wg tezaurusa na dysku. Jest to tzw. indeksowanie nadmiarowe dokumentów w EMC.

d) gramatyka opisu deskryptorowego: brak wskaźników roli (ważności), natomiast są wskaźniki więzi np.

1 DeskprA, Deskr.B .....

2 Deskr.C, Deskr.D .....

e) rola tezaurusa w systemie - podczas wprowadzania dokumentów programy systemu INIS zapewniają kontrolę terminologiczną, korektę błędów i hierarchiczne oznaczanie deskryptorów w dokumentach wejściowych.

f) wyszukiwanie - brak danych o metodzie wyszukiwania.

II. EURATOM - Nuclear Science Abstracts.

a,b,c,d,e,f - jak wyżej (INIS)

g) dane dotakowe - 60-70.000 dokumentów rocznie. Obecnie zbiór posiada 500 tys.dok. a całkowity czas wyszukiwania 65 pytań wynosi około 18 min.

III. GOLEM - Gra speicherorientiertiv Listenstenorganisierte Ermittlungsmethode.

a) system-wyszukiwawczo-wydawniczy (najnowocześniejszy na świecie opracowywany na Olimpiadę 1972)

b) typ EMC - Siemens System 4004/45 (pamięci dyskowe, bębnowe i kartowe magnetyczne - brak taśm magn.)

c) indeksowanie dokumentu - współrzędne, deskryptorami s tezaurusa,

d) gramatyka opisu deskryptorowego - brak wskaźników więzi (np. zdań deskryptorowych w jednym dokumencie) ale jest zaznaczenie wskaźników roli poprzez utworzenie jednostek indeksowania

e) rola tezaurusa w systemie - tezaurus może zawierać 300 tys. deskryptorów - brak dokładnych danych o hierarchii tezaurusa. W tezaurusie są podane częstotliwości użycia danego deskryptora w dokumentach

f) wyszukiwanie - przy pomocy kartotek inwersyjnych zorganizowanych na dysku (brak danych o budowie tego zbioru). Kryterium wyszukiwania - 10 deskryptorów. Operatory logiczne - operator dysfunkcji, koniunkcji i negacji. Deskryptory, które wystąpiły w pytaniu, a nie ma ich w tezaurusie, są sygnalizowane np. deskryptora A - brak w tezaurusie. Jeden przebieg wyszukiwania trwa max. 30 sek. przy zbiorze podstawowym dokumentów około 1 mln oraz w tezaurusie o 300.tys. deskryptorów przy wyszukiwaniu odpowiedzi na pytania zawierające 10 deskryptorów.

IV. IRMS - Information Retrieval and Management System

a) system wyszukiwawczy ogólnego zastosowania (niekoniecznie informacyjny)

b) typ maszyny - IBM-360. Wymagany system dyskowy DOS, programy w języku podstawowym tzn. asembler

c) indeksowania dokumentów - współrzędne, deskryptorami

d) gramatyka opisu deskryptorowego - brak gramatyki tzn. indeksowanie deskryptorami z tezaurusa bez nadawania cech ważności (tzn. roli) i bez zaznaczenia zdań (więzi). Średnia ilość deskryptorów w dokumencie wynosi 10.

e) rola tezaurusa w systemie - tezaurus zawiera relacje synonimiczności, hierarchii oraz kojarzenia a także przydzielone dziedziny semantyczne w poszczególnych deskryptorach. System tezaurusa zorganizowany jest na dysku.

Uwaga: system formowy IRMS pozwala również na tworzenie i uaktualnianie tezaurusa w EMC. Jednak brak dokładnych danych na temat organizacji tego zbioru.

f) wyszukiwanie - przez kartoteki inwersyjne zorganizowane na dysku magnetycznym. Wyszukiwanie zarówno na pytania retrospektywne jak i profilowe.

Operatory logiczne - jak w systemie GOLEM tzn. dyzjunkcji koniunkcji i negacji.

V. FIND 2 - File Interrogation of Nineteen Hundred Data.

a) system wyszukiwawczy ogólnego zastosowania (podobnie jak (IRMS)) system firmowy f-my ICL.

b) typ EMC - seria ICL-1900

c) indeksowanie dokumentów - przy pomocy określonego słownika, który także jest w maszynie. Ponieważ ten system wyszukiwawczy nie był przeznaczony do wyszukiwania informacji bibliograficznej - brak jest zbioru tezaurusa z hierarchią, jest tylko słownik np. wszystkich danych personalnych lub zapasów magazynowych itp. w EMC.

d) gramatyka opisu deskryptorowego - brak wskaźników roli i więzi, indeksowanie współrzędne, słowami ze słownika

e) rola tezaurusa - brak

f) wyszukiwanie - bardzo ciekawa technika wyszukiwania wykorzystująca zbiory taśmowe - tworzenie "łańcuchów" i zbioru trafień. Oprócz utworzenia zbioru podstawowego (głównego) dokumentów system tworzy serię podzbiorów opartych na kombinacji i permutacji danych ze zbioru głównego, są to tzw. łańcuchy. Zbiór ten jest tak zorganizowany na taśmie, że tylko wymaga jednokrotnego przeglądnięcia zbiorów i odszukania dokumentów na całą serię pytań.

VI. MISTRAL - Compagnie Internationale pour l'Informatique - CII

a) system wyszukiwawczo-wydawniczy informacyjny

b) typ EMC - IRIS-50 (f-ma franc.)

c) indeksowanie dokumentów: - współrzędne, deskryptorami z tezaurusa) szczegółowymi.

Natomiast podczas wprowadzania do EMC wszystkie deskryptory opisu deskryptorowego dokumentu są automatycznie uzupełniane deskryptorami hierarchicznie szerszymi a także synonimami. Jest to jedyny spotkany w literaturze przypadek tego typu indeksowania nadmiarowego dokumentów.

d) gramatyka opisu - brak danych

e) rola tezaurusa - zbiór tezaurusa zorganizowany na dysku zawiera relacje synonimii i hierarchii. Tezaurus zawiera 60 tys.deskr., z których każdemu może odpowiadać aż 100 pojęć nadrzędnych

f) wyszukiwanie - na pytania retrospektywne jak i periodyczne. Operatory logiczne - do formułowania pytań używa się trzech operatorów logicznych, tj. dyzjunkcji, koniunkcji i negacji.

w stosunku do deskryptorów treściowych oraz relacje mniejszości, większości i równości dla danych numerycznych. Maksymalna liczba pytań profilowych w systemie wynosi 60 tys., w trakcie jednego przetwarzania można grupować 50 pytań.

VII. System MEDLARS (Narodowej Biblioteki Medycznej USA) wykorzystujący tezaurusa i opisy deskryptorowe dokumentów

VIII. System ASCA (Automatic Subject Citation Alert-ISI) wykorzystujący wykaz słów w opisach bibliograficznych.

IX. System STAIRS (Storage and Information Retrieval System)
   a) typ maszyny - IBM-360
   b) system konwersacyjny
   c) umożliwia wyszukiwanie retrospektywne jak również profilowe

X. System PASSAT (Program zur Automatischen Selektion von Stichworten aus Texten)
   a) typ maszyny - SIEMENS-4004
   b) umożliwia automatyczne indeksowanie dokumentów
   c) słowa kluczowe posiadają wagi

XI. System SMART (Salton's Magical Automatic Retriever of Texts)
   a) typ maszyny - IBM
   b) klasyczny system wyszukiwania informacji tekstowej w języku angielskim, przeznaczony do przebadania różnych algorytmów analizy tekstu i wyszukiwania informacji na żądanie
   c) indeksowanie dokumentu - automatyczne
   d) algorytm automatycznej analizy tekstu wykorzystuje wagi oraz współczynniki korelacji poszczególnych słów tekstu

a) wyszukiwanie informacji - istnieje możliwość zmniejszenia czasu wyszukiwania informacji poprzez ograniczenie przeszukiwań do pewnej grupy dokumentów zebranych uprzednio w tzw. "wiązki"

XII. System SOCCER (Smart's own Coordance Constructor, Extremely Rapid)

a) typ maszyny - IBM-360

b) umożliwia szybkie tworzenie słownika dla analizowanego tekstu z wykazaniem dla każdego słowa jego położenia w tekście. Jest to tzw. konkordancja tekstu.

XIII. System NEW YORK TIMES

a) typ maszyny IBM-360

b) system wyszukiwawczy

c) wprowadzane są do maszyny pełne wycinki informacji z gazet (w języku naturalnym)

d) jest efektywny dla małych objętościowo informacji. Dokumentowanie publikacji zwartych (książek, czasopism) jest niemożliwe

XIV. System ALTAIR (Automatic Logical Translation and Information Retrieval)

a) typ maszyny - CDC-3400 oraz IBM

b) faktograficzny system wyszukiwania w dziedzinie astronomii

c) zbiór informacji dotyczy 9-ciu tysięcy systemów gwiezdnych

d) algorytm wyszukiwania - powolny. Czas wyszukiwania jest proporcjonalny do liczby terminów użytych w formule wyszukiwania

XV. System MARC (Machine - Readable - Catalog) - system Biblioteki Kongresu USA

a) typ maszyny - IBM

b) wydawniczy, określenie formatu danych bibliograficznych na maszynowych nośnikach informacji.

Załącznik 3

Systemy wyszukiwania informacji oparte na Uniwersalnej Klasyfikacji Dziesiętnej

I. System ILACS (Integrated Library Administration and Cataloging System) - Holandia, 1969 r.
   a) typ maszyny cyfrowej - IBM-360/40 (od 1971 - IBM 360/50)
   b) język informacyjny - UKD oraz słowa kluczowe znaczące z tytułów
   c) dziedzina zastosowań - chemia i technologia żywności
   d) w systemie tym szczególny nacisk położono na organizację różnego rodzaju kartotek.

II. System GIBUS (Groupe Informatique de Biblioteque Universitaire et Scientifique) - Francja
   a) typ maszyny cyfrowej - IBM-360/40,
   b) język informacyjny - UKD
   c) dziedzina zastosowań - międzydyscyplinarny
   d) jest to system wyszukiwawczy mający elementy systemu konwersacyjnego.

III. System AUDACIOUS (Automatic Direct Access to Information with On-line UKD System) - USA, 1968 r.
   a) typ maszyny cyfrowej - IBM-7044
   b) język informacyjny - UKD oraz tezaurus EURATOMu
   c) dziedzina zastosowań - nukleonika
   d) jest to pierwszy system, w którym zastosowano UKD do konwersacyjnego systemu wyszukiwania. Eksperyment ten przeprowadzili Freman i Artelhon na zbiorze 3 tysięcy dokumentów.

IV. System CFSS (Combined File Search System) - USA

a) typ maszyny cyfrowej - IBM-1401

d) język informacyjny - UKD oraz hasła przedmiotowe

c) dziedzina zastosowań - oceanografia

d) w pytaniach wuzględniono operatory logiczne: "i", "lub", "negacja". System wykorzystuje kartotekę liniową dokumentów oraz kartoteki inwersyjne słów kluczowych.

V. System AMCOS (Aldermaston Mechanised Cataloging and Orders System)

a) typ maszyny cyfrowej - IBM-360/50

b) język informacyjny - UKD

c) dziedzina zastosowań - nukleonika

d) zastosowano w tym systemie cykliczne przetwarzanie informacji.