



RAZVOJ IN UVEDBA RAČUNALNIŠKEGA SISTEMA ZA PREGLEDOVANJE PRISPEVKOV V MEDIJIH

Andrej Prevc, HermesPlus d.d., Ljubljana, Šlandrova 2
Gorazd Kogoj, ICOS d.o.o., Ljubljana, Dunajska 51

Povzetek

V prispevku predstavljamo svoje izkušnje pri uvedbi sistema za elektronski zajem, distribucijo, analizo in shranjevanje novinarskih prispevkov iz različnih medijev. Pri tem dajemo poseben poudarek shranjevanju podatkov v binarnih objektih relacijske baze, ki sicer služi za delo s podatki o dokumentih. Uporaba orodij za hiter razvoj uporabniških vmesnikov ima pomembno vlogo pri uspešni uvedbi sistema in njegovi prilagodljivosti spremembam, prav tako kot ustrezna metodologija uvajanja.

1. UVOD

Prispevek je nastal kot posledica teoretičnih in praktičnih spoznanj, do katerih smo prišli pri realizaciji projekta "InfoKlip" (elektronsko arhiviranje časopisnih člankov in magnetogramov) za Urad vlade za informiranje kot naročnika in Centra vlade za informatiko kot investitorja. Osvetlili bi radi nekatere pomembne elemente tega projekta, predvsem pa uporabo relacijske baze in prednosti, ki nam jih le ta prinaša. Služi naj kot orientacija in vzpodbuda vsem tistim, ki bodo v svojem okolju razvijali podobne sisteme za arhiviranje in se bodo srečali s podobnimi problemi kot avtorji tega prispevka. Glede na to, da je bil projekt ocenjen od naročnika, investitorja in izvajalca kot zelo uspešen, lahko trdimo, da smo dosegli vse načrtovane cilje in to s pomočjo domačega znanja in razvoja.

2. IZHODIŠČA IN CILJI

Vsakomur se je verjetno že zgodilo, da je prebral kakšno vest, izjavo, intervju, gledal TV prispevek in si čez čas zaželel, da bi ta prispevek ponovno prebral, si ga ponovno ogledal. V kopici časopisov in revij, ki jih imamo vsak dan pred seboj, pa se pogosto zgodi, da prispevka, ki bi ga želeli ponovno prebrati, enostavno ne najdemo več.

Na Uradu vlade za informiranje so se s takimi primeri redno srečevali. Pogosto so dobili zahtevo naj najdejo prispevke o neki temi, ki so bili objavljeni v različnih medijih v daljšem časovnem obdobju. Za taka iskanja so porabili tudi po več dni in vendar na koncu ni bilo mogoče reči, da je bilo delo res v celoti opravljeno. Podobno je bilo pri povpraševanjih v slogu: "Članek o temi to pa to, ki je bil objavljen v Delu ali pa v Dnevniku pred kakšnim mesecem". Pogosto je bilo iskanje celo neuspešno. Poleg tega

je eden od bistvenih pogojev za uspešno delo vlade, ministrstev in drugih državnih organov, da so sprotno in celovito seznanjeni s tem, kaj o njihovem delu meni javnost.

To so bili bistveni razlogi za to, da so na Uradu vlade za informiranje začeli razmišljati o vzpostavitvi urejene zbirke prispevkov, ki se nanašajo na vso paleto področij, ki jih pokrivajo vladni resorji. Urejeni bi morali biti po ključnih besedah in drugih atributih, ki omogočajo preprosto in hitro iskanje prispevkov.

Tako so skupaj s Centrom vlade za informatiko konec leta 1994 zasnovali projekt "InfoKlip".

Poglejmo, kaj so uporabniki pričakovali da bodo pridobili s sistemom InfoKlip. Osnovne funkcije, ki naj bi jih omogočil sistem, so:

- zajem - optično čitanje dokumentov (slike člankov)
- shranjevanje optično včitanih dokumentov (slike)
- zajem in shranjevanje besedil v različnih izvornih oblikah (urejevalniki besedil, ASCII teksti ipd.)
- opremljanje zajetih dokumentov (člankov) s podatki (atributi), ki pozneje omogočajo hitro iskanje
- uporabniški vmesnik v slovenščini
- v elektronski in tiskani obliki
- integriran sistem za statistično vrednotenje in prikazovanje podatkov v grafični in tabelarni obliki
- hitra iskanja po vnešenih atributih, ključnih besedah in prostemu besedilu ter kombinaciji vseh teh elementov.

Poleg teh splošnih zahtev, ki jih lahko v veliki meri izpolni marsikateri programski paket, ki je na razpolago na tržišču, pa so bile postavljene še dodatne zahteve, ki so bistveno vplivale na našo strategijo:



- hiter vnos in obdelava velike količine podatkov o člankih s strogimi časovnimi omejitvami
- vgradnja orodja za prepoznavanje besedila (OCR) v isti uporabniški vmesnik za vnos podatkov o dokumentih
- integracija ostalih informacijskih virov v enoten sistem (STA, sporočila,)
- delo v heterogenem komunikacijskem okolju
- možnost razširitve strojne opreme (upsizing) brez spreminjanja programske opreme
- preprosta uporaba brez daljšega šolanja uporabnikov
- hiter, zanesljiv sistem v skladu s konceptom odjemalca-strežnika
- možnost dostopa do vseh podatkov tudi z oddaljenih delovnih mest.

Sami pa smo pri izbiri upoštevali še naslednje bistvene smernice:

- možnost za nadaljni razvoj in razširitev programske opreme na željo uporabnika (naročnika)
- varnost sistema in preprosto vzdrževanje.

Na javnem razpisu smo kandidirali z rešitvijo, ki jo opisujemo v nadaljevanju.

3. KAKO POISKATI ENOTNO REŠITEV ZA VSE NAROČNIKOVE ZAHTEVE?

Glede na opisane zahteve in usmeritve, ki jih je postavil naročnik, smo se odločili, da bomo realizirali celotni sistem z lastnim znanjem in razvojem. Zakaj? Odgovor je zelo preprost - na tržišču pač ne obstaja že izdelan produkt, ki bi v celoti izpolnjeval vse postavljene zahteve.

Tehnološke opredelitve, ki smo si jih zastavili kot razvijalci, so bile: relacijska baza podatkov, okolje Unix kot srce sistema, okolje MS Windows za uporabnike, standardi (I250, 852, Tiff G4, TCP/IP), strojno in programsko komprimiranje ter dekomprimiranje optično čitanih dokumentov, koncept odjemalca-strežnika.

Relacijska baza podatkov

Samo sistem, ki temelji na relacijski bazi podatkov, omogoča odprt in standardiziran dostop do shranjenih podatkov. Vsi ostali načini predpostavljajo poznavanje strukture in načina shranjevanja podatkov in so dostopni samo tistim aplikacijam, ki poznajo te strukture.

Večina sistemov za arhiviranje dokumentov uporablja relacijske baze le za shranjevanje alfanumeričnih podatkov, ne pa za shranjevanje dokumentov samih. Ti so ponavadi shranjeni kot del zunanjih datotečnih sistemov. Tak koncept pomeni dva različna načina shranjevanja podatkov, enega za alfanumerične podatke, drugega za dokumente. V tem primeru bi morali razviti in vzdrževati

dva različna sistema, pojavlja pa se nam tudi vprašanje integritete in varnosti podatkov celotnega sistema. Shranjene dokumente bi lahko spreminjali ali brisali mimo relacijske baze, v kateri so shranjeni osnovni podatki o dokumentih. Težava, ki bi nam jo lahko povzročal tak sistem, je tudi izdelava varnostnih kopij, ki bi jih morali izvajati na dva različna načina. Dodatno pa bi se lahko stvari zapletle ob uporabi eksternih optičnih diskov (jue-boxov), ki omogočajo še tretji način zapisa podatkov. Še en zelo pomemben problem, ki se pojavi pri uporabi takega koncepta, je kako zagotoviti enoten prenos podatkov od strežnika do odjemalca in obratno v LAN in WAN okolju. Orodja za izdelavo aplikacij (odjemalca-strežnika) omogočajo le delo s podatki, ki so shranjeni v podatkovni bazi, ne pa s samostojno shranjenimi datotekami na strani strežnika ali odjemalca. V takih primerih je treba razviti še dodaten komunikacijski protokol ter aplikacije in to za vsako platformo posebej.

Zato smo izbrali drugačen koncept - to je shranjevanje dokumentov in alfanumeričnih podatkov v enotno bazo na podatkovno komunikacijskem strežniku z operacijskim sistemom Unix. To nam omogočajo moderne relacijske baze, v katere lahko shranjujemo in obdelujemo nekonvencionalne podatkovne tipe (tekstovne zapise, slike, zvok itd.). Pri uporabi takega koncepta seveda odpadejo vse prej omenjene težave in omejitve. Tako smo se lahko bolj posvetili vsebinskemu delu celotnega projekta, kar je bilo glede na časovno omejitve (4 mesece), izredno pomembno za uspešnost projekta.

V okoljih s heterogenimi sistemi je postavitve sistema na osnovi koncepta RDBMS nujnost. Zapiranje v produkte, ki ne podpirajo SQL pristopa, pomeni ukinitve modela "odjemalca-strežnika". Tudi performančno, ko gre za velike količine podatkov, so relacijske baze edina izbira.

Glede na praktične izkušnje, ki smo jih imeli z relacijskimi bazami, ter performančne kazalce, smo se odločili za "Informix OnLine" z dodatkom "Informix Online Optical". Slednji nam omogoča uporabo optomagnetnih diskov, ki tako postanejo del enotne baze.

Izdelava uporabniškega okolja

Pri razvoju grafičnih aplikacij, ki delujejo v okolju MS/Windows, smo se odločili za razvojno orodje Informix HyperScript Tools, ki ima direktni dostop do relacijske baze Informix OnLine brez vmesnika ODBC. S tem orodjem je možno hitro in učinkovito razviti grafične uporabniške vmesnike ter jih integrirati z eksternimi programi, ki so izdelani v programskem jeziku C++. Orodje omogoča tudi direkten prikaz statističnih podatkov v obliki grafov in tabel, kar je bila ena od pomembnejših zahtev naročnikov. Uporabili smo tudi programsko orodje C++, z katerim smo razvili posamezne module za krmljenje



optičnih čitalcev (scanner) in module za hitro dekomprimiranje in prikazovanje optično včitanih dokumentov. Poleg tega pa smo v prikazovanje vgradili tudi optično prepoznavanje (OCR) celotnega besedila ali pa samo izbranega dela dokumenta. Vse to smo integrirali v enoten uporabniški vmesnik, ki je enostaven in prijazen za uporabnika, obenem pa izvaja vse funkcije, ki jih je naročnik zahteval. Pripravili smo več vrst uporabniških vmesnikov, odvisno od funkcij, ki jih mora vsebovati (optično včitavanje in obdelava, samo obdelava in distribucija, razširjena analitična obdelava, uporaba sistema od končnih zunanjih uporabnikov itd.).

Procesi, ki potekajo v ozadju sistema

Pri tako kompleksnem sistemu kar nekaj procesov poteka tudi v ozadju celotnega sistema, ki jih uporabnik nikoli ne opazi, za delovanje celotnega sistema pa so zelo pomembni. Vsi ti procesi se izvajajo na podatkovnem strežniku Unix občasno ali pa ves čas delovanja sistema. Naj navedemo samo nekaj najbolj pomembnih: avtomatski sprejem novic slovenske tiskovne agencije (STA) in vpisovanje le-teh v podatkovno bazo, avtomatsko tiskanje velikih količin optično včitanih dokumentov - člankov za distribucijo, indeksiranje tekstovnih prispevkov, avtomatska izdelava varnostnih kopij, itd. Moduli so bili napisani v programskem jeziku C z direktnimi ukazi za delo s podatkovno bazo. To nam je omogočil Informix-ov produkt "Embedded SQL" za programski jezik C.

4. IMPLEMENTACIJA

Za uspešno uvajanje sistema za delo z dokumenti morata biti izpolnjena naslednja pogoja:

- izvajalec, ki pripravlja sistem, mora poznati problematiko, ki jo bo sistem reševal,
- uporabniški vmesniki morajo biti narejeni tako, da sledijo načinu razmišljanja ljudi, ki delajo z dokumenti; to so običajno uporabniki, ki jih računalništvo sploh ne zanima in dojemajo računalnik zgolj kot orodje (kar je prav).

Postopek uvajanja uporabniškega vmesnika za vnos podatkov o člankih smo izvedli po preprosti osnovni metodologiji. Zagotoviti smo morali, da bo logika dela vnašalcev prilagojena njihovim miselnim tokovom in bo sistem na ta način čim bolj povečal hitrost in natančnost vnosa.

Da bi ugotovili, kakšen način dela je najboljši za vnašalce, smo najprej pripravili dva postopka vnosa na osnovi specifikacije naročnika. Potem smo povabili potencialne vnašalce na preizkus, ki je trajal več dni po nekaj ur, pri tem pa smo zapisali vsako pripombo in predlog, ki je bila pri poskusnem delu omenjena. Tako smo na primer bistveno spremenili zaporedje podatkov, ki jih je treba

vnesti, obliko vnosa za nekatere podatke in podobno. Tak uporabniški vmesnik vsebuje samo tiste funkcije, ki so potrebne v delovnem procesu, ne pa dodatnih možnosti, ki bi uporabnika pri delu bolj motile kot pa mu pomagale.

Ena od posledic takega postopka pri razvoju je tudi to, da nov uporabnik sistema ne potrebuje usposabljanja v pomenu seznanjanja s sistemom izbir in načinov vnosa ampak samo vsebinsko usposabljanje glede smiselno pravičnega vnosa podatkov.

Tudi za izdelavo vmesnika za brskanje po arhivu smo uporabili podobno metodo. Osnovno izvedbo smo uporabili kot predlog na podlagi katerega so bodoči uporabniki ugotovili, kaj jim ustreza in kaj jih moti. Pri zasnovi tega vmesnika smo imeli ves čas v mislih tudi optimizacijo uporabe zvez, saj je bila neposredna uporaba sistema z oddaljenih delovnih mest ena od osnovnih zahtev.

Tehnološke osnove, za katere smo se odločili in so opisane v prejšnjem poglavju, so se v fazi udejanjanja izkazale kot ustrezne, saj smo lahko realizirali ves podane zahteve in brez težav vgradili vse dodatne funkcije, ki jih je dodatno želel naročnik.

Rezultat takega sprotne prilagajanja sistema in vmesnikov ljudem, ki ga uporabljajo, je uspešno in hitro delo. Poleg tega pa so uporabniki, ki so sodelovali pri oblikovanju sistema, dosti bolj motivirani za uspešno uvedbo sistema. To se je pozneje pri neposredni uporabi sistema tudi izkazalo, saj je le-ta neprimerno bolj izkoriščen kot sistemi, ki jih uporabniki razumejo kot nekaj vsiljenega.

Na koncu bomo navedli nekaj dejstev, ki ilustrirajo prednosti opisane tehnologije.

Sistem je bil uveden v treh fazah:

- Izdelava pilotske verzije
- Testiranje pilotske verzije in zbiranje pripomb
- Uvedba končne verzije v uporabo

Za te tri faze smo potrebovali pet mesecev dela z aktivnim sodelovanjem uporabnikov v preizkusih. V uporabniški vmesnik, ki sicer omogoča optično čitanje tiskanih člankov in zajem besedil prispevkov v elektronski obliki, smo vgradili tudi orodje za optično prepoznavanje besedila v delu članka, ki ga vnašalec sam določi.

Sistem zdaj deluje brez prekinitev že skoraj dve leti. V tem času je bilo vnešenih pribl. 135.000 slik člankov, ki se shranjujejo na optičnih diskih WORM. Poleg tega je v bazi že več kot 300.000 prispevkov, ki izvirajo bodisi iz servisa STA ali pa so zapisi intervjujev in prispevkov na televizijo oziroma radiju. Vse te podatke vključno z vsebinami prispevkov (slike ali besedila) neposredno krmilimo z orodji relacijske baze, kar nam tudi zagotavlja nadzorovan dostop in varnost. Vnos 250 do 400 člankov in njihovo označevanje se izvede vsak dan v manj kot treh urah, s čimer je izpolnjen eden od bistvenih ciljev projekta - hiter vnos.



Statistična obdelava podatkov o člankih je avtomatska, njen rezultat pa so preglednice in grafi, ki jih uporabnik dobi ob pritisku določenega gumba. Ti pokažejo trende in rezultate dela z javnostjo in mediji.

V sistem so bili po tem vgrajeni še prispevki in podatki iz drugih virov, ki jih zajemamo na različne načine (zunanji dobavitelji tovrstnih informacij, elektronska distribucija prek modemov, ipd.), česar ni bilo v prvotni zahtevi.

Uveden je bil tudi uporabniški vmesnik za pregledovanje člankov in iskanje po bazi z oddaljenih delovnih mest, pri čemer je bila uporabljena komunikacijska hrbtnica Centra vlade za informatiko za državno upravo. S pomočjo tega vmesnika lahko vsak uporabnik sistema poišče prispevke na osnovi izbirnih podatkov, ki jih pozna. Pri tem ima na razpolago spiske možnih vrednosti, kar bistveno olajša iskanje in zmanjša možnost neprijetnih napak pri sprotnem vnosu iskalnih besed. Uporaba takega vmesnika je zelo preprosta in posebno uvajanje sploh ni potrebno, ker uporabnik takoj vidi, katere možnosti ima na razpolago. Običajno s strani uporabnikov ni bilo vprašanj potem, ko so bili seznanjeni z načinom dela. Ta uvod pa traja od 30 minut do 45 minut.

Na ta način je bilo pregledovanje podatkov in iskanje predano v roke uporabnikom, ki lahko sproti preverijo ustreznost svojih iskanj in si najdene prispevke iztiskajo na svojem delovnem mestu.

Koncept rešitve je prav pri delu z oddaljenih lokacij dokazal svojo učinkovitost, saj je delo prek zvez minimalno upočasnjeno v primerjavi z delom v lokalni mreži,

kljub zahtevnosti prenosa slik dokumentov. Preizkusili pa smo tudi delo prek običajnih modemskih povezav, ki pa je v skladu s pričakovanji počasnejše.

Produkcija teče brez intervencij tehnične službe investitorja, prav tako za vzdrževanje sistema ni potreben posebej izšolan človek. Redna opravila so namreč avtomatizirana ali pa toliko poenostavljena, da jih lahko opravi tudi laik. Bistveno je tudi to, da sistem deluje zanesljivo in brez napak.

5. SKLEPNE MISLI

Tehnološke osnove in orodja, ki smo jih uporabili, omogočajo maksimalno prilagodljivost sistema naročnikovim zahtevam. Glede na zelo specifične zahteve, kot na primer celovita obdelava podatkov z izračuni trendov in podobnimi statističnimi pregledi, naknadne uvedbe zajemanja podatkov iz novih virov (kar je naročnik zahteval nakladno, ko je sistem že deloval) brez temeljne reorganizacije sistema ipd., je bila naša izbira gotovo upravičena. Vsi standardni izdelki, ki jih poznamo in so dosegljivi na trgu, lahko izpolnjujejo velik del zahtev. Da pa bi izpolnili še zadnje, običajno porabimo bistveno več sredstev in časa, kot smo ga potrebovali mi za izdelavo paketa od začetka.

Vsi cilji, ki so bili zastavljeni na začetku, so bili do sedaj doseženi. Prav tako pa smo pripravljeni na prilagoditve, ko bo prišlo do novih zahtev.