

## PROBLEMI PRI SHRANJEVANJU PODATKOV - - VEČ PODATKOV VEČ PROSTORA

Miroslav Milovanovič\*

UDK: 004.3:930.25

Miroslav Milovanovič: *Problemi pri shranjevanju podatkov - več podatkov več prostora. Tehnični in vsebinski problemi klasičnega in elektronskega arhiviranja. Zbornik referatov z dopolnilnega izobraževanja, Maribor 5/2006, št. 1, str. 389-394.*

Izvirnik v slovenščini, izvleček v slovenščini in angleščini, povzetek v angleščini.

Vsakodnevno se srečujemo s problemi, kako učinkovito shranjevati digitalizirane dokumente, kjer pa moramo biti pozorni tudi na probleme, ki nas še čakajo v prihodnosti in kako se prilagoditi takšnim spremembam. Rešitve za kompleksnost takšnega procesa se sprejemajo vsak dan, zato imamo možnost preizkušanja, katera je najbolj primerna za implementacijo v poslovne sisteme, s katerimi bomo delovali skozi digitalizacijo.

UDC: 004.3:930.25

Miroslav Milovanovič: *Problems with the preservation of information – more information, more space. Technical and Field Related Problems of Traditional and Electronic Archiving. Conference Proceedings, Maribor 5/2006, No. 1, pp. 389-394.*

Original in Slovenian, abstract in Slovenian and English, summary in English.

In modern days we are faced with the problem of how to save digitalized documents efficiently, taking into consideration potential problems and possible solutions to such problems. Solutions to the complexity of the process are adopted every day, so we can test which are the most suitable for the business in which we are trying to implement such a dynamic informational procedure.

### UVOD

Elektronsko arhiviranje je področje, ki z nenehnim spreminjanjem in izboljševanjem, vedno znova skrbi za preglednejše in racionalno poslovanje. V zadnjih letih smo priča pospešeni digitalizaciji arhivskih gradiv. Digitalizacija pa se ne ustavi samo na papirnih dokumentih, ampak se vse aktivneje digitalno arhivirajo tudi druge analogne oblike (avdio, video zapisi ...).

Pri organizaciji se pri procesu arhiviranja znova pojavljajo novi izzivi, kako neko klasično arhivsko gradivo učinkovito pretvoriti v digitalno obliko in kako ga v bodoče shranjevati na način, da je dostopen naročniku. Od vseh področij pri digitalnem arhiviranju med ostalimi izstopa tudi prostorska ureditev, s katero se sooča vse več izvajalcev, ki opravljajo digitalizacijo arhivov.

Vedno večje število dokumentov pa tudi vse zahtevnejši pogoji zajema narekujejo dimenzijo digitalne oblike, kar zahteva tudi več prostora za shranjevanje. Reševanje problemov se usmerja predvsem na iskanje razmerja med velikostjo ter kakovostjo končne rešitve.

---

\* Miroslav Milovanovič, informatik organizator, Organizacija informatike, Osenjakova 4, 1000 Ljubljana, Slovenija.

## RAZVOJ TEHNOLOGIJE ZA ARHIVIRANJE

Tehnologija, ki se uporablja znotraj digitalizacije arhivov, je po večini različna od materiala, ki se arhivira. Zaradi prilagoditve in optimizacije poslovanja pa podjetja za digitalizacijo večinoma razvijajo tudi lastne programske rešitve.

V podjetju MFC&L smo zaradi vse večjega pogojevanja naročnikov uvedli lastno programsko rešitev ADS (Arhivsko dokumentarni sistem), ki omogoča hitro prilagoditev zahtevam naročnikov. Komerzialne različice arhivskih programov nam niso omogočale hitrega prilagajanja in natančnosti pri delu, skozi lastni razvoj pa lahko prilagodimo storitve tako, da ustrezajo naročnikom.

Pri tehnologiji zajema je smiselno izpostaviti naslednje oblike digitalizacije, ki zagotavljajo kakovost izhodnih dokumentov:

1. OCR (Optical character recognition)
  - o Pri OCR tehnologiji se dokument iz klasično grafičnega zajema digitalno obdela tako, da sistem opravi prepoznavo znakov in posreduje kot izhodno obliko dokument v tekstovnem formatu.
2. OMR (Optical Mark Recognition)
  - o Pri OMR tehnologiji sistem pri npr. skeniranju odčitava vrednost oziroma simbole ali oznake na za to pripravljenih obrazcih (prijavne pole na vzgojno izobraževalnih ustanovah, ankete ...). Za prepoznavo se uporablja posamezen del dokumenta (okvirji). Prepoznavo pri OMR tehnologiji ima samo dve izhodni vrednosti, in sicer ali je izbrano ali ni. Znotraj konfiguracije delovnih modelov pa se lahko prilagodi sistem tako, da zagotavlja dodatno analizo podatkov (statistika) in ne samo indeksno vrednost. OMR zagotavlja predvsem konsistentno obdelavo podatkov na osnovi standardov ali oblik, ki so sprejeti znotraj organizacije poslovanja. Kot primer OMR tehnologije je najbolj razširjena oblika prepoznavanja črtnih kod, ki je tudi ena od najstarejših oblik zapisovanja identifikacijskih podatkov.
3. ICR (Intelligent Character Recognition)
  - o ICR tehnologija je tehnologija, ki je še v razvoju, ponuja pa poleg standardnega prevajanja simbolov tudi prevajanje pisave oziroma ročnega pisanja. Projekt, čigar dimenzija je vsekakor rešitev za vse, ki jim iz kateregakoli vzroka ni mogoče uporabljati sodobne tehnološke rešitve.

Pri zgoraj omenjenih metodah pretvarjanja podatkov pri prostorskem shranjevanju ne naletimo na večje težave, saj je razmerje velikost slike/tekstovna datoteka (XML, tekstovne datoteke) precejšnje in se s tem tudi olajša odločitev za končno obliko.

4. Slikovni zajem
  - o Pri slikovnem zajemu je mišljena predvsem digitalizacija klasičnega arhiva, lahko kot pretvorba iz digitalnega tekstovnega formata ali pa zajem podatkov s skeniranjem.

Prostorsko se razlika kaže v tem, da je konverzija pri digitalnih podatkih racionalnejša, saj je pri zajemu podatkov s skeniranjem veliko dejavnikov (sence, nepotrebni podatki, ločljivost ...), ki vplivajo na velikost izhodne datoteke.

Slikovni zajem je eden od najbolj razširjenih formatov pri digitalizaciji klasičnih arhivov, posledično pa zaradi vse večjih zahtev po izboljšanju kakovosti tudi prostorsko obremenjujoč.

#### 5. VR (Voice Recognition)

- o V zadnjih letih smo priča razvoju tehnologije za prepoznavo govora in s tem tudi avtomatizem pri konverziji iz avdio formata v tekstovno obliko. Ponudniki aplikativnih rešitev so redki (Dragon Naturally Speaking ...), predvsem pa je pomanjkljiva jezikovna podpora. Tovrstne rešitve vsekakor sodijo bolj med optimizacijo poslovanja, vendar ne gre zanemariti učinek konverzije iz avdio v tekstovno obliko, kjer se kaže razlika v porabi prostora in preglednost pri arhiviranju.

### KAKOVOST ARHIVSKEGA GRADIVA IN KAKOVOST ZAJEMA (STANJE ORIGINALOV IN REZULTAT ZAJEMA)

Proces zajema dokumentarnega gradiva se začne s pregledom dokumentacije, od katerega je predvsem odvisna hitrost, velikost in natančnost zajema. Sam pregled je različen od tehnologije, ki je uporabljena za zajem dokumentov. Medtem ko nekateri organizacijsko urejajo materijo, tako da je pred zajemom urejena in ima posledično v digitalni obliki linearno in pregledno pot, pa drugi zajemajo dokumente brez potrebnih ključev za urejanje in jih določajo naknadno s pomočjo indeksnih atributov.

Stanje arhivske materije je ključno za kakovosten zajem in kakovostno digitalno obliko. Ocena izvedljivosti zajema v digitalno obliko je predpogoj za prilagajanje nadaljnjega procesa na vseh področjih, tako z vidika strojne opreme kot programske rešitve ter organizacije dela. Dokumentom v stanju, ki onemogoča normalen zajem, je potrebno zagotoviti obliko, ki ne ovira delovnega procesa, predvsem pa preglednost zajema z ostalimi podatki ter njihovo skladnost.

Kakovost zajema je rezultat digitalizacije dokumentov, njihova izhodna oblika pa je različna od pogojev, ki so postavljeni pred samo izvedbo.

Sama kakovost zajema je opredeljena predvsem z dvema pogojeva:

- Tehnični pogoji zajema (Znotraj tehničnih pogojev naročnik opredeli kakšno obliko mora imeti dokument, ko gre skozi proces zajema, npr. skeniranje dokumentov na določeni ločljivosti, format datoteke, kompresija ...).
- Tehnologija, ki se uporablja pri zajemu dokumentarnega gradiva (Tehnologija je tisti element, ki zagotavlja konkurenčno prednost, saj imamo že sedaj možnost spremljati razvoj izdelkov, ki v večini podpostopkov pri digitalnem arhiviranju ponujajo celovito rešitev).

Vsekakor zgoraj navedena pogoja nista edina, ki vplivata na končno obliko digitalnega dokumentarnega gradiva, saj je pri tako zahtevnem postopku od začetka odvisno tudi, v kakšnem stanju je gradivo, ko je podvrženo digitalizaciji. Za večino dokumentacije bi lahko rekli, da je izhodna oblika enaka vhodni, oziroma če je dokument v slabem stanju, potem ni mogoče zagotoviti pomanjkljivih delov. V nekaterih primerih je dovoljena manipulacija podatkov zaradi zagotavljanja kakovosti (ta manipulacija se nanaša predvsem na dokumentacijo, pri kateri je tudi v naročnikovem interesu odprava napak, npr. odstranjevanje šumov pri avdio zapisih).

## PROBLEMI PRI SHRANJEVANJU (PROSTORSKI IN OSTALI)

Problemi pri shranjevanju niso vezani samo na sam prostor, temveč tudi na način, kako omogočiti vpogled v dokumente in hitrost pregledovanja samega dokumenta.

Živimo v dobi globalnih omrežij, ko se z razvojem tehnologije ponuja tudi oddaljen dostop do gradiva, ki ga hočemo videti.

Cilj je zagotoviti vpogled v dokumentacijo ali arhiv na način, da bo lahko naročnik iz oddaljene lokacije hitro in pregledno brskal po vsebini arhiva. Problemi, ki se kažejo pri shranjevanju dokumentov, so vedno vezani na prostor, kjer se shranjuje.

Če arhiviramo npr. skenirane dokumente v A0 velikosti in je bila zahteva naročnika, da ne sme biti uporabljena kompresija pri shranjevanju, lahko ena takšna datoteka doseže 500MB velikosti in več (ločljivost 400 dpi). Če gre pri projektu za majhno število takšnih dokumentov, potem se lahko arhivska materija uredi z običajnimi prostorskim rešitvami (cd, dvd mediji, diskovna polja ...). Problem nastane, ko je zahteva naročnika, da je potrebno originalno obliko datotek hraniti trajno in hkrati omogočati hitro pregledovanje npr. slik, velikosti formata A0. Ko govorimo o nekaj tisoč takšnih datotekah, potem govorimo tudi o nekaj TB prostora za shranjevanje le teh.

Problem, ki se kaže pri zgoraj omenjenem procesu, mogoče niti ni toliko pereč z vidika zagotovitve prostora, problem nastane, ko moramo do teh datotek preko elektronskega arhiva in še posebej preko omrežij omogočiti dostop in vpogled. Samo arhiviranje je lahko problematično tudi zaradi vzdrževanja, v primeru, da imamo sisteme, ki nam ne dovolijo spreminjanja ali brisanja podatkov. Če arhiviramo podatke, za katere nimamo zakonskih določil glede trajanja, je pa zahteva naročnika, da se uporabi tehnologija, ki onemogoča brisanje ali spreminjanje podatkov, potem je prostorsko še toliko bolj rizična hramba takšnih podatkov, saj ni zagotovila, da bo naročnik toliko časa, za kolikor so bili arhivirani podatki, tudi operativen.

## PROSTORSKE REŠITVE

Prostorske rešitve so različne od namena in predvidene uporabe digitaliziranih dokumentov. Kot že omenjeno, je treba opredeliti, ali bodo dokumenti samo hranjeni v elektronski obliki ali pa gre v procesu digitalizacije za servisne storitve, s katerimi bomo omogočali vpogled v te digitalne dokumente.

Poznamo več rešitev, s katerimi rešujemo tovrstne težave:

- *izločanje nepotrebne, neuporabne, neveljavne, podvojene ali zakonsko dopustno uničljive dokumentacije iz klasičnega in digitalnega arhiva*

V postopku digitalizacije je prva rešitev preverjanje materiala, ki je namenjen pretvarjanju v elektronski format. S tem se izognemo nepotrebni podvajanju neuporabnih in zakonsko zastarelih dokumentov.

- *sistemske rešitve, ki omogočajo enkratnost zapisov (rešitev, ki je mogoče pomembnejša za preglednost digitalnih dokumentov, s tem pa se zmanjšuje dodajanje identifikacijskih elementov)*

- *digitalni formati z velikim faktorjem stisljivosti rastrskih datotek, ki ohranijo visoko kakovost podatkov (lossless compression), npr. DjVu format.*

Stisljivost datotek je pomemben del digitalizacije, saj tukaj kot končni del procesa določamo velikost datoteke glede na kakovost kompresije. Pomembnost se kaže tudi v zagotavljanju majhnosti zaradi dostopa preko omrežij (za ogled grafične datoteke, ki je velika tiff - 500MB ali pa Jpeg - 80MB, predstavlja problem prenosa tako velike količine podatkov preko npr. spleta). Ena od rešitev, ki je bila razvita predvsem za ogledovanje grafičnih datotek preko spleta, je DjVu kompresija oziroma DjVu format. Format, ki podpira izjemno močno kompresijo z majhno izgubo podatkov (lossless compression), postaja prednostna oblika pri prikazovanju grafičnih datotek, čigar izvirna velikost presega standardno velikost grafičnih datotek. DjVu format zagotavlja do 100 % stisljivost z izjemno majhno izgubo podatkov (npr. tiff 300MB v DjVu 3MB). Seveda je potrebno pri tem upoštevati navodila naročnika, ki določa, v kakšnem formatu mora biti digitalizirana materija.

- *izbira formata, v katerega se bo pretvarjalo dokumentarno gradivo.*

Izbira formata je prav tako pomemben korak pri digitalizaciji. Omejitve, ki jih ima izbira formata, so predvsem pravno pogojene oziroma določene s standardi saj le-te določajo način in obliko, s katero neka organizacija posluje (npr. če hočemo, da je digitalni dokument skladen z poslovanjem nekega drugega sistema, moramo tudi zagotoviti prepoznavnost tega dokumenta na drugem sistemu).

Rešitev, kot je zgoraj omenjeno, je več. Vsaka organizacija prireja rešitve glede na obstoječo informacijsko infrastrukturo in vizijo poslovanja.

## ZAKLJUČEK

Glede na zahteve po kakovosti zajetih ali ustvarjenih digitalnih dokumentov in z rastjo števila poslovne dokumentacije se večja tudi potreba po prostoru, ki naj bi varno shranjeval to dokumentacijo. Informacijska tehnologija nam z razvojem nudi nekatere rešitve, globalno pa se rešitve še razvijajo.

## SUMMARY

### PROBLEMS WITH PRESERVATION OF INFORMATION - - MORE INFORMATION, MORE SPACE

The article points to certain problems within the digitalizing process. We can divide them into the following topics:

Informational technology: here the emphasis is on already defined ways of digitalizing documents and processes that are used.

Quality of archives: assessment of the condition of archives in the process of the digitalization with regard to the fulfilment of the customers' needs.

Problems with saving: the actual problems in the current informational structure and the potential problems.

Solutions for efficient saving: some solutions that are used to prevent the waste of space and some solutions for optimizing business process.

*Miroslav Milovanovič je trenutno še absolvent na Fakulteti za upravo, kjer pripravlja diplomsko nalogo z naslovom E-volitve. V podjetju MFC&L je zaposlen od 1. 1. 2006 kot informatik organizator, kjer je njegova naloga raziskava in razvoj informacijskega področja znotraj podjetja in prilagajanje informacijskim spremembam.*