OD DIGITALNE MINLJIVOSTI DO DIGITALNE TRAJNOSTI

Jacqueline Slats [*]

UDK: 004.3:930.253

***Jacqueline Slats: Od digitalne minljivosti do digitalne trajnosti**. Tehnični in vsebinski problemi klasičnega in elektronskega arhiviranja. Zbornik referatov z dopolnilnega izobraževanja, Maribor 4/2005, št. 1, str. 145-155.*

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

Glede na nizozemski arhivski zakon in predpise se arhivsko gradivo predaja dvajset let po nastanku v "dobrem stanju, urejeno in dostopno". Za digitalno gradivo pomeni dvajset let več kot je njegova življenjska doba. Zaradi tega je Nizozemski državni arhiv skupaj z Ministrstvom za notranje zadeve oktobra 2000 pričel z raziskovalnim projektom digitalne hrambe (Digital Preservation Testbed), s pomočjo katerega bi zagotovili predpise za primeren pristop k varstvu digitalnega gradiva.

DIGITALNI SPOMIN DRŽAVNE UPRAVE

Digitalna državna uprava se je zdela v prejšnjem stoletju tako zelo oddaljena, sedaj, v 21. stoletju, pa uprava vedno bolj in bolj posluje z digitalnim gradivom. Komunikacija s pomočjo elektronske pošte je postala del vsakodnevne rutine, elektronske podatkovne zbirke pa uporabljajo vsepovsod. Državna uprava je zato zavezana obravnavati te informacije na ustrezen način.

Digitalno gradivo je potrebno ustrezno hraniti, saj mora ostati dostopno bodočim generacijam. Ta princip velja tudi za papirno gradivo, s katerim upravljamo in ga hranimo. Gradnja digitalne državne uprave zahteva čim prejšnjo namestitev primerne digitalne infrastrukture. Gradiva ni potrebno samo hitro najti, ampak mora biti tudi avtentično in berljivo (ne glede na trenutno tehnologijo) in takšno ostati tudi v prihodnosti.

Sedanja nizozemska vlada namerava do leta 2006 izvesti 65 % vseh postopkov med državno upravo in državljani digitalno. Leta 2002 je bil cilj 25 % in je bil z lahkoto dosežen. Zaradi tega imajo trenutno veliko dela na področju razvoja strategij, metod, tehnik in orodij za odgovorno vodenje digitalnih postopkov državne uprave.

DIGITALNA HRAMBA

Najpomembnejša težava, ki se nanaša na hrambo izvirnega digitalnega gradiva je tehnološka zastarelost. Tehnološke spremembe so večje iz dneva v dan. To poraja mnogo vprašanj, npr. kaj storiti z dokumenti, ki so bili ustvarjeni s staro strojno in programsko opremo, ki je ni mogoče več uporabljati. Če ne bomo ukrepali sedaj, nimamo zagotovila, da bomo dokumente, ki nastajajo danes, lahko brali v prihodnosti s prihodnjo tehnologijo.

[*] ***Jacqueline Slats, Head Digital Longevity, Nationaal Archief of the Netherlands.***

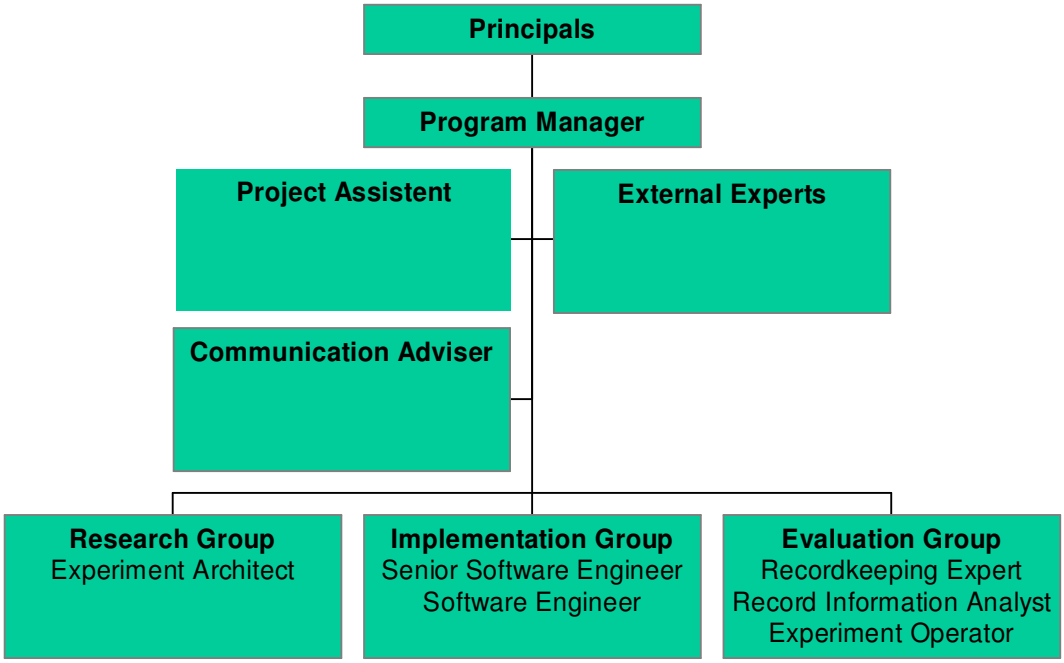
RAZISKOVALNI PROJEKT DIGITALNE HRAMBE (Digital Preservation Testbed)

Raziskovalni projekt sta oktobra 2000 ustanovila Ministrstvo za notranje zadeve in Ministrstvo za izobraževanje, kulturo in znanost (Nizozemski državni arhiv pa je organ v sestavi ministrstva). To je triletni raziskovalni projekt, katerega cilj je bil raziskati možnosti zagotavljanja trajnega dostopa do avtentičnega arhivskega gradiva skozi daljše časovno obdobje.

Projekt je bil praktičen raziskovalni projekt, v okviru katerega smo izvajali poskuse v nadzorovanem in varnem okolju. Na podlagi tega smo lahko preverili učinke in posledice izvedenega varstva na arhivskem gradivu. Usmerjala so nas posamezna raziskovalna vprašanja, ki smo si jih zastavili ob pričetku projekta.

RAZISKOVALNA SKUPINA

Zastavljen pristop je zahteval multidisciplinarno skupino, ki so jo sestavljali strokovnjaki s področja informacijsko-komunikacijskih tehnologij, arhivisti, domači in tuji strokovnjaki in drugi. Zelo dragocena je bila testna skupina za ocenjevanje, ki ni omenjena v spodnjem diagramu, sestavljali pa so jo arhivisti iz različnih institucij npr. Državnega arhiva Nizozemske, arhivskega inšpektorata in davčne službe. Vladne institucije, ki so nam zagotavljale kopije gradiva, so sodelovale v skupini med potekom eksperimentov.



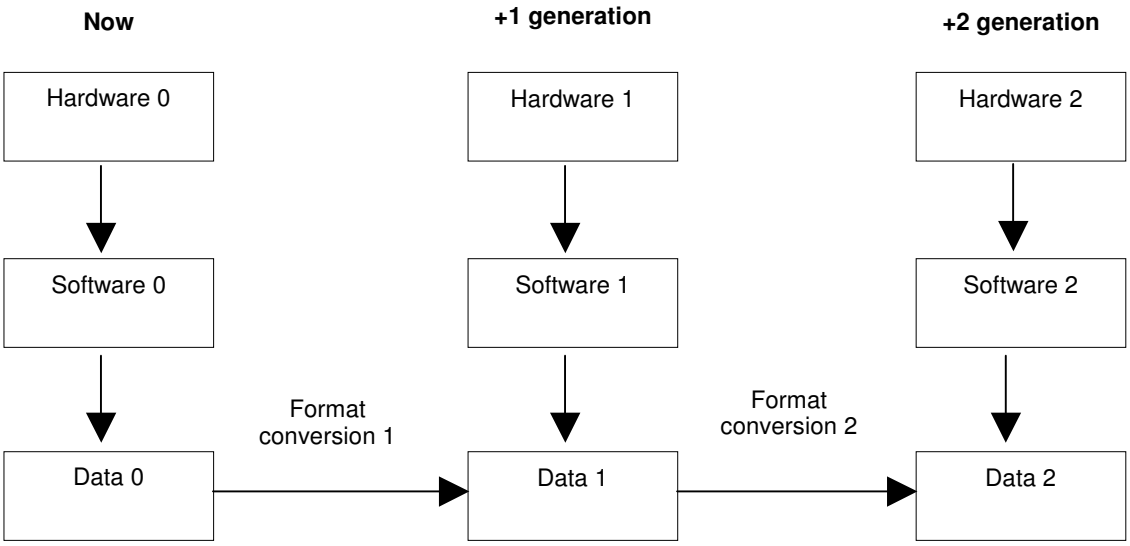
Shema organizacije projekta

PRISTOPI K VAROVANJU DIGITALNEGA GRADIVA

Predmet raziskav so bili trije različni načini pristopa k dolgodobnemu digitalnemu varovanju: »migracija«, »XML« in »emulacija«. Ocenjevali pa niso zgolj posameznih pristopov, temveč tudi njihove meje oz. zmožnosti, stroške in možnosti uporabe.

MIGRACIJA

Obstaja več vrst definicij za migracijo. V projektu smo migracijo definirali kot konverzijo zapisa iz ene strojne ali programske opreme v drugo.



Osnovni diagram migracije

V projektu smo proučevali in eksperimentirali z naslednjimi oblikami migracije:

- združljivost s predhodnimi okolji,
- interoperabilnost,
- konverzija v standarde.

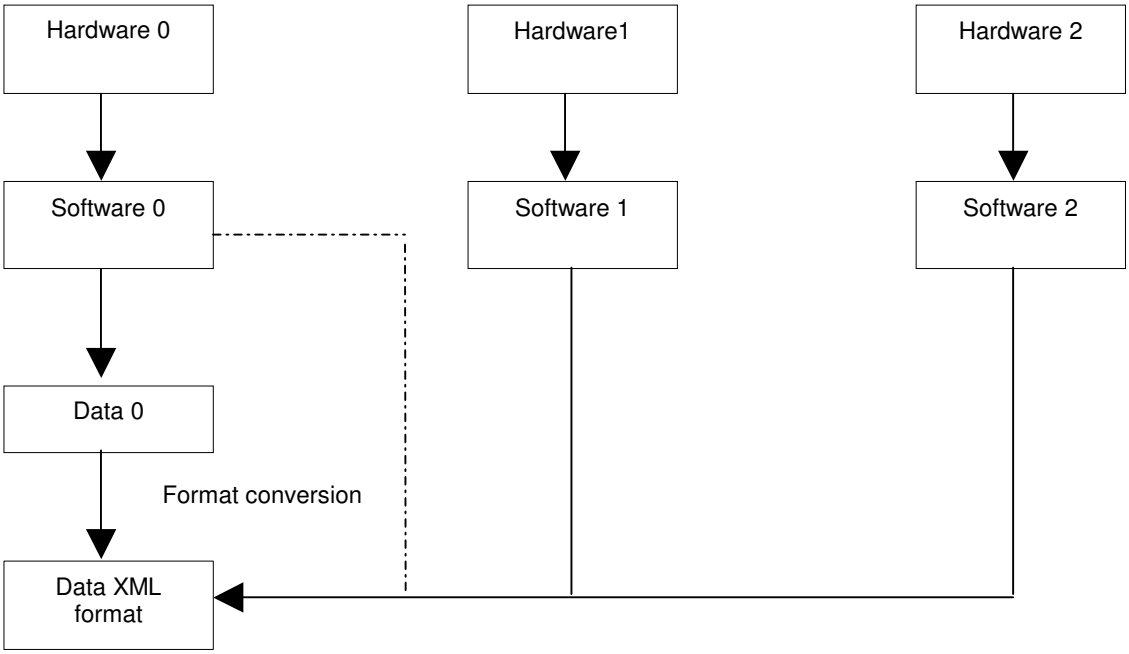
XML

V projektu smo kot možnost pristopa k dolgoročnemu varovanju digitalnih zapisov preučevali tudi razširljivi označevalni jezik ali XML, kar je kratica za eXtensible Mark-up Language. To je poseben označevalni jezik, ki omogoča obogatitev podatkov z informacijami o njihovi strukturi, kar pomeni, da ga je mogoče uporabljati kot datotečni format. Gre za odprt standard, ki ga je postavila neprofitna organizacija World Wide Web Consortium, v kateri razvijajo interoperabilno tehnologijo, kot so specifikacije, navodila, programska oprema in orodja, ki omogočajo uporabo interneta kot celote.

XML ni odvisen od specifične platforme in ga lahko z uporabo enostavnega urejevalnika teksta berejo tako ljudje kot tudi naprave. Zaradi tega lahko XML uporabljamo za digitalno hrambo. Odvisno od načina uvedbe XML pristopa, se le-ta lahko prekriva z drugimi, zgoraj opisanimi strategijami. Npr. konverzijo datotek v XML lahko smatramo za specifično obliko migracije (glej zgoraj: konverzija v standarde).

XML je zasnovan tako, da omogoča računalniškim programom enostavno obdelavo, kar je tudi eden od razlogov, da predstavlja dober format za hrambo; v prihodnosti bo relativno lahko napisati programsko opremo za obdelavo danes ustvarjenih XML datotek.

Datoteke lahko konvertiramo v XML neposredno ali pa jih neposredno generiramo v XML datotečnem formatu. Ker XML ni odvisen od določene kombinacije strojne in programske opreme, je bolj vzdržljiv kot mnogi komercialni datotečni formati. Tako se bo zmanjšalo število konverzij, kakor tudi tveganje za ogrožanje avtentičnosti digitalnih zapisov.



Konverzija v XML format zahteva manj konverzij kot migracija

UNIVERZALNI VIRTUALNI RAČUNALNIK

Pristop na podlagi emulacije, ki uporablja univerzalni virtualni računalnik ali UVC (Universal Virtual Computer) se nekoliko razlikuje od izvirnega koncepta emulacije. Emulator je kljub temu potrebno napisati, vendar v tem primeru za neobstoječ, virtualni računalnik, t. i. UVC. UVC je računalnik z enostavno zgradbo in osnovnim nizom navodil, ki bi jih v prihodnosti moral znati napisati vsak razvijalec programske opreme. UVC se nato uporabi za zagon aplikacije (UVC dekoder podatkovnega formata), ki vzame kot vhodne zapise izvirne dokumente, izhodni zapis pa je v obliki logičnega podatkovnega opisa ali LDD (Logical Data Description). Ta

logični podatkovni opis je sestavljen iz oznak, ki zagotavljajo dodatne informacije o vsebini digitalnih dokumentov. Dodatne semantične informacije so sestavljene tako, da bi morali biti v prihodnosti ljudje sposobni interpretirati logične podatkovne opise brez dodatnih pripomočkov. Tako bo pregledovalnik, izdelan v prihodnosti, obdelal logični podatkovni opis, ki bo nato pokazal avtentičen digitalni dokument na zaslonu.

UVC strategija hrambe se samo delno opira na emulacijo in vsebuje tudi nekatere vidike migracijske strategije. Pri uporabi UVC-ja konvertiramo izvorne datoteke s pomočjo programa, napisanega v UVC programskem jeziku, v LDD. LDD je neodvisen, samo opisen in jasno strukturiran podatkovni format, z informacijami, ki bodo v prihodnosti potrebne za ponovno sestavo digitalnih zapisov.

UVC HRAMBA PODATKOV

"Hramba podatkov" je prva in najosnovnejša izvedbena oblika UVC strategije. V njej se podatki – izvorna datoteka in izvorni format – shranijo s programom, ki ekstrahira podatke iz bitnega niza in jih enostavno in neodvisno opiše, tako, da lahko pregledovalnik obdela podatke.

Izvorna datoteka, npr. JPEG datoteka, je shranjena skupaj s specifičnim UVC programom za dekodiranje podatkovnih formatov za JPEG. V prihodnosti bo ta UVC JPEG program tekel na UVC emulatorju. UVC JPEG program prebere bitni niz izvorne datoteke in izdela izhodni zapis v obliki LDD. LDD se reproducira na bodoči računalniški platformi z uporabo pregledovalnika, ki ga bodo lahko razvili v prihodnosti na podlagi LDD sheme.

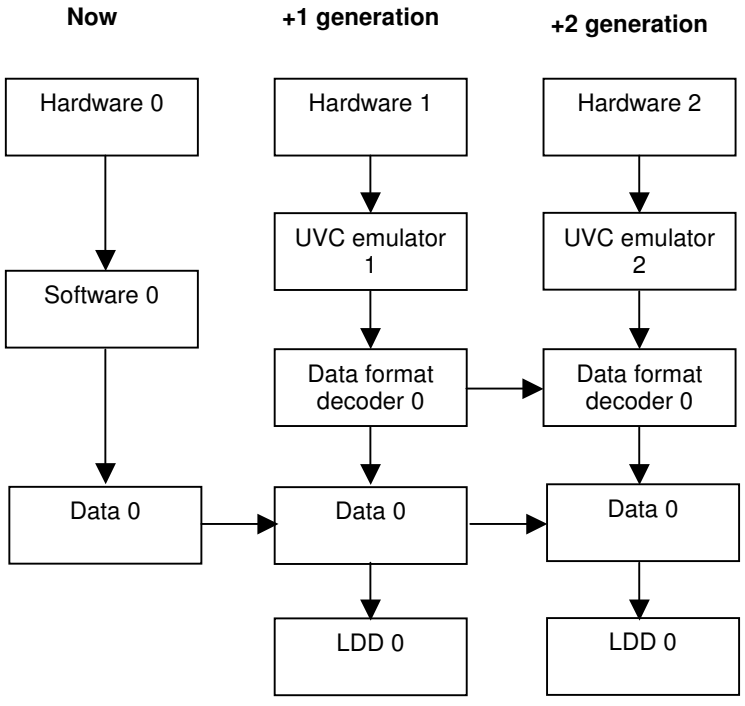


Diagram Univerzalnega virtualnega računalnika (Universal Virtual Computer)

Izvorni bitni niz je pri tej strategiji nespremenjen in nova datoteka (LDD), ki nastane, ko zaženemo UVC program za dekodiranje podatkovnih formatov, ni shranjena. LDD se pokaže s pomočjo pregledovalnika. Format in struktura LDD sta tako natančno definirana, da bi izdelava novega pregledovalnika morala biti enostavna. Če bo potrebno, se lahko razvijejo novi pregledovalniki za bodoče računalniške platforme.

Trenutno je za vsak tip LDD-ja potreben poseben pregledovalnik. To pomeni, da je verjetno potrebno uporabiti stotine različnih pregledovalnikov. V sledeči fazi razvoja UVC-ja, bodo oblikovane posamezne kategorije objektov, ki delujejo s podobno logiko. Posamezna kategorija takšnih objektov, npr. datoteke v različnih slikovnih formatih, bo ustvarila en LDD, za katerega bo potrebno razviti en pregledovalnik. Vseeno pa bo še vedno potrebno razviti posamezen UVC program za dekodiranje podatkovnih formatov za vsakega od le-teh.

Pomanjkljivost UVC emulacijskega pristopa je, da morajo biti UVC programi za dekodiranje podatkovnih formatov napisani za vsak posamezen tip datotek (za generiranje opisa logičnih podatkov). Poleg tega pa je potrebno napisati nove UVC emulatorje za vsako novo generacijo programske opreme, ki se tako razlikuje od prejšnje, da star UVC emulator na njej ne more več delovati zanesljivo.

Če bomo želeli, da postane UVC izvedljiva in delujoča strategija za dolgoročno hrambo digitalnih zapisov, bo potrebno glede na mnogo različnih vrst formatov in tipov digitalnih zapisov, razviti celo vrsto programov za dekodiranje. Dokončen uspeh UVC strategije pa je deloma odvisen tudi od tega, v kolikšni meri bodo to strategijo sprejeli proizvajalci strojne in programske opreme. Proizvajalci programske opreme bodo morali razviti UVC programe za dekodiranje podatkovnih formatov za svojo programsko opremo, da bo omogočen logičen podatkovni opis, ki bo temeljil na izvornih datotekah. Ko bo prišlo do tega, se bo lahko UVC strategija zelo razširila.

EKSPERIMENTI

Preizkuse smo izvajali na štirih različnih vrstah zapisov: tekstovni zapisi, tabele, e-pošta in na podatkovne zbirke različnih velikosti, zapletenosti in narave. Gre za tipe zapisov, ki jih v 90 % uporablja nizozemska državna uprava.

Elektronske zapise smo klasificirali glede na pet lastnosti, ki jih je identificiral Rothenberg. To so: vsebina, kontekst, struktura, videz/pojavnost in obnašanje.

RAZISKOVALNA VPRAŠANJA

Zastavljena vprašanja so razdeljena na tri poglavitna področja: splošno, metapodatki in lastnosti. Splošna vprašanja vključujejo:

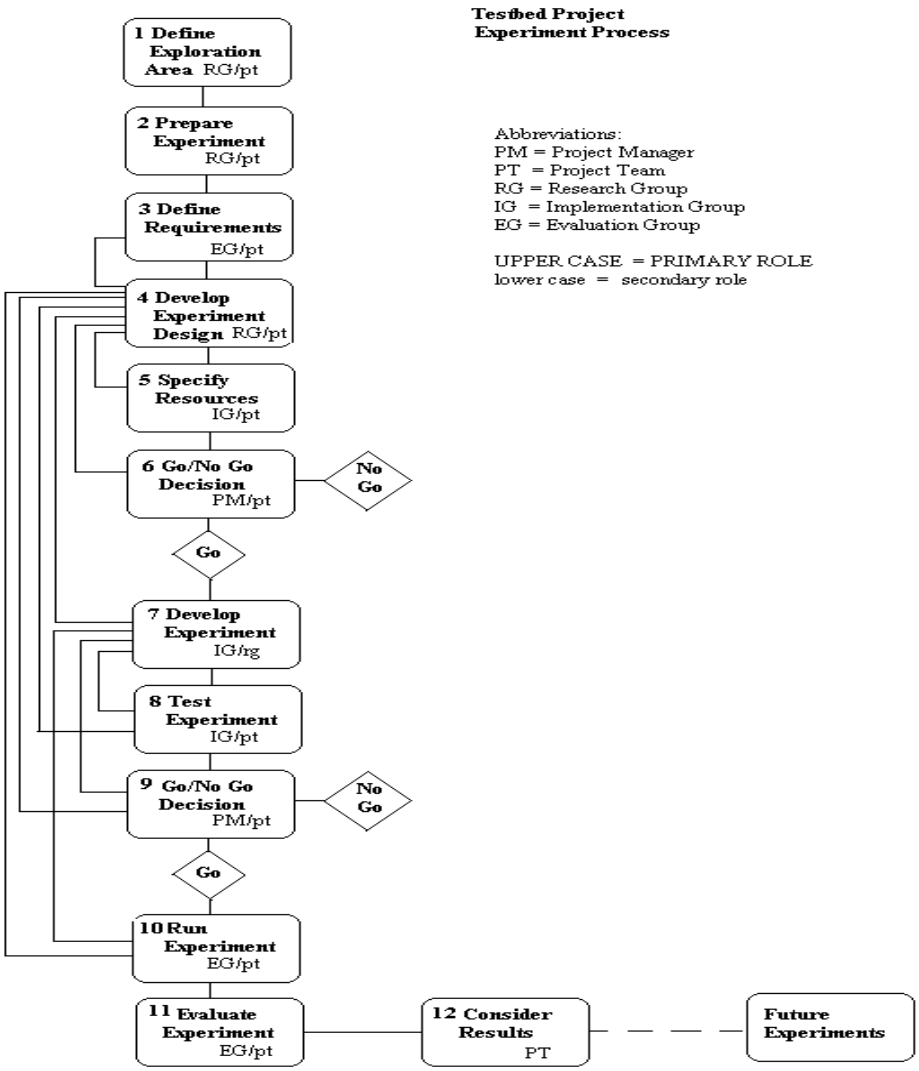
- Kaj so prednosti in slabosti uvedbe posameznega načina hrambe?
- Kako lahko merimo oziroma demonstriramo učinkovitost in uspešnost posameznega načina?
- Kateri faktorji vplivajo na učinkovitost in primernost vsakega posameznega načina npr. stroški, tipi zapisov, zahteve po avtentičnosti, roki hrambe?
- Katere so osnovne zahteve funkcije hrambe, npr. kakšne so zahteve za dostop in ponovno uporabo (retreival) zapisov po hrambi?

Vprašanja o metapodatkih vključujejo:

- Kateri faktorji vplivajo na metapodatke, potrebne za hrambo npr. tip zapisov in način hrambe, in kako?
- Kakšne so opcije združevanja metapodatkov z zapisi?

EKSPERIMENTALNI POSTOPEK

Da bi lahko nadzorovali projekt in izvajali poizkuse v varovanem okolju, smo razvili dvanajststopenjski eksperimentalni proces. Tu smo tudi jasno določili, predvsem s študijem razpoložljivih publikacij, če je kateri tip zapisa izključen iz katerega izmed načinov hrambe. Ti koraki so natančno dokumentirani v podatkovni zbirki projekta. V času eksperimenta smo nadzorovali zapise, da bi ugotovili, ali in kako so specifične metode primerne za dolgoročno hrambo.



Eksperimentalni postopek

REZULTATI

ELEKTRONSKA POŠTA

Pri elektronski pošti smo se odločili, da bomo eksperimentirali samo z XML načinom. Na osnovi predhodnega študija se je izkazalo, da je elektronska pošta posebej primeren tip zapisa za XML. Obstaja veliko podobnosti med XML formatom in formatom zapisa elektronske pošte, tako da je konverzija med obema enostavna. Oba sta jasno določena.

Elektronska pošta mora slediti t. i. internetnemu sporočilnemu formatu (Intenet Message Format), da je lahko interoperabilna na različnih platformah. Ta format je dobro zasnovan in definira sestavne dele osnovne datoteke prenosa elektronske pošte (basic email transmission file). Trenutni standard, ki se uporablja za elektronsko pošto je RFC 2822, z ekstenzijo MIME, specificirano v RFC 2045–2049. Nadzoruje ga neprofitna organizacija Internet Engineering Task Force. Standard je dobro definiran in strukturiran ter temelji na čistem besedilu.

XML je standardiziran format tako kot jezik za označevanje (mark-up). Prav tako je dobro določen, nadzira pa ga neprofitna organizacija World Wide Web Consortium. W3C je odgovoren za organizacijo in vzdrževanje XML specifikacij, sheme, standarda in XSLT priporočil. XML je, kot kaže že ime samo, razširljiv. Lahko ga priredimo in razširimo za kateri koli namen, pa vendar ostane sam sebi zvest. Deluje lahko na kateri koli strojni ali programski platformi in ga lahko beremo na katerem koli urejevalniku navadnega besedila.

Podrobnosti med obema formatoma pomenita, da je postopek konverzije relativno enostaven. Vsi posamezni deli so enostavno označeni v datoteki poštnega prenosa (email transmission file) in jih je enostavno spremeniti v podobno dobro strukturirano XML datoteko.

Obstajata dve različni možnosti konverzije v XML:

- »po uporabi« (naknadno konvertiranje v XML) in
- »pred uporabo« (neposredno generiranje v XML).

»Po uporabi« je namenjen že obstoječi pošti (tako izhodna kot prispela sporočila), ki jo je potrebno hraniti za nedefiniran čas (zaradi tega se ta sporočila konvertirajo kasneje).

»Pred uporabo« pa se lahko uporablja za nova izhodna sporočila in je prvi korak v smeri trajne hrambe uradne elektronske pošte (sporočila se neposredno ob svojem nastanku generirajo v XML).

TABELE

Pri tabelah smo se odločili, da bomo eksperimentirali z vsemi tremi načini. Posebno velik izziv je bilo eksperimentiranje z UVC načinom hrambe, kajti tabele imajo več nivojev (npr.: podatkovni nivo, nivo formularjev).

Čeprav je koncept UVC-ja izredno obetajoč, se je generiranje logičnega podatkovnega opisa izkazalo za težavno. Do tega ni prišlo zaradi kompleksnosti UVC-ja, ampak zaradi pomanjkanja potrebne dokumentacije o datotečnih formatih. Iz poročila Nizozemske nacionalne knjižnice smo razbrali, da so naleteli na isto težavo.

Migracija zapisov iz stare verzije v novo verzijo iste aplikacije, npr. Excel 97 v Excel 2000, je uporabna za kratkotrajno hrambo. Rezultati tega preizkusa so bili primerljivi z rezultati migracije tekstovnega dokumenta v novejšo verzijo.

Kot primeren format za avtentično predstavitev tabel, vključno z različnimi sloji, se je izkazal XML.

TEKSTOVNI DOKUMENTI

Kot prva načina eksperimentiranja s tekstovnimi dokumenti smo izbrali migracijo in XML. Za UVC način smo uporabili poročila Nizozemske nacionalne knjižnice, v katerih smo dobili dokaze, da je koncept primeren za shranjevanje elektronskih publikacij.

Migracija zapisov iz starejše verzije v novejšo verzijo iste aplikacije, npr. Word 97 v Word 2000, je uporabna zgolj za kratkotrajno hrambo. Pri konvertiranju teh zapisov v sodobnejše verzije nismo naleteli na pomembne težave. Presenetljivo je bilo, da so rezultati bili še boljši, če smo preskočili eno ali več verzij. Vendar pa lahko po večkratni konverziji večje ali manjše spremembe vplivajo na avtentičnost zapisa. Zaradi tega je potreben ročni nadzor. Še več, migracijo je potrebno ponavljati vsakih nekaj let, to pa je izvedljivo le, če je migracija avtomatizirana.

Za migracijo tekstovnih zapisov v standardni format smo eksperimentirali z PDF-jem in RTF-jem. PDF je primeren za predstavitev tekstovnega dokumenta avtentično, posebej glede vsebine in oblike.

Migrirali smo tudi stare zapise, ustvarjene na enem urejevalniku besedila v drugega, npr. WP 4.2 v Word 2002. Pri tem načinu smo zahteve v zvezi z avtentičnostjo dosegli le po ročni intervenciji.

Pri načinu XML smo ugotovili, da lahko zagotovi avtentičnost konteksta, vsebine, strukture in obnašanja izvornega tekstovnega dokumenta. Za predstavitev izgleda pa je potrebna dodatna tabela.

PODATKOVNE ZBIRKE

Pri eksperimentiranju s digitalnimi podatkovnimi zbirkami smo se soočili z vprašanjem: "Kaj je arhivsko gradivo?"

- celotni sistem podatkovne zbirke (podatkovna zbirka, DBMS in uporabniška aplikacija),
- podatkovna zbirka kot takšna,
- vrsta/niz v tabeli podatkovne zbirke,
- zapis, sestavljen iz polj, razširjenih po različnih tabelah ali
- podatki iz podatkovne zbirke, pridobljeni in predstavljeni na natančen način, kot so bili v aplikaciji.

Kljub opravljenim predhodnim raziskavam in diskusijam z arhivskimi strokovnjaki, nismo mogli najti nedvoumnega odgovora na to vprašanje. Zato smo se s pragmatičnega stališča odločili, da bomo eksperimentirali s celotnim sistemom podatkovne zbirke in s podatkovno zbirko kot tako.

Migracija podatkovne zbirke iz starejše v novejšo verzijo iste podatkovne zbirke, npr. Access 97 v Access 2000, je uporabna za kratkotrajno predstavitev konteksta, vsebine, oblike, strukture in obnašanja. Rezultati so primerljivi z rezultati migracije tekstovnih dokumentov in tabel v novejše verzije.

Konverzija v XML je primerna za predstavitev konteksta, vsebine in strukture podatkovne zbirke kot takšne. Da bi ohranili tudi obliko aplikacije, je dodatno potrebo shraniti tudi tehnično in funkcionalno dokumentacijo o sistemu podatkovne zbirke, vključno z zajemom zaslonskih slik (screen shots).

Z uporabo migracije oz. XML-a nismo mogli ohraniti obnašanja sistema podatkovne zbirke za daljše obdobje. Prav tako tega nismo mogli doseči z UVC načinom. Potencialni način bi lahko bila emulacija strojne opreme, vendar le-te nismo vključili v arhivske vidike.

PERSPEKTIVA

Nizozemski državni arhiv je za naslednja leta definiral tri nove projekte:

- priporočila, ki se nanašajo na navodila in predpise o arhiviranju,
- priporočila projekta, vključena v sistem Elektronskih dokumentov in upravljanja z dokumenti ministrstev, in
- projekt emulacije strojne opreme.

Podrobnejše informacije na
spletni strani: http://www.digitaleduurzaamheid.nl,
e-pošta: Testbed@nationaalarchief.nl.

SUMMARY

FROM DIGITAL VOLATILITY TO DIGITAL PERMANENCE

Digital Preservation Testbed was a three-year practical research project with the overall goal of investigating options to secure sustained accessibility to authentic archival records over the long-term, by carrying out experiments in a controlled and secure environment. This allowed us to ascertain the effects of undertaken preservation action on archival records.

Testbed was researching three different approaches to long-term digital preservation: migration, XML and emulation. Not only the effectiveness of each approach was evaluated, but also their limits, costs and application potential.

Experiments took place on four different record types: text documents, spreadsheets, emails and databases of different size, complexity and nature.

At the end of 2003 the Digital Preservation Testbed project provided:

- Advice on how to deal with current digital records
- Recommendations for an appropriate preservation approach or a combination of approaches per record type
- Functional requirements for a preservation function
- Cost models of the various preservation strategies
- A decision model to select the right preservation strategy
- Recommendations concerning archival guidelines and regulations

Prevod v slovenski jezik: Zdenka Semlič Rajh