

Μια ασυμβίβαστη αρχιτεκτονική για την συντήρηση ψηφιακών εγγράφων

Thomas A. Phelps and P.B. Watry

University of Liverpool

Liverpool, Great Britain

phelps@ACM.org, P.B.Watry@liverpool.ac.uk

Περίληψη.

Το πολλαπλών χρήσεων (πολυχρηστικό) πρότυπο εγγράφου προσφέρει μια πρακτική, αποδεδειγμένη, ασυμβίβαστη αρχιτεκτονική για να συντηρεί ψηφιακά έγγραφα ενδεχομένως οποιουδήποτε μορφότυπου δεδομένων. Εφαρμόζονται ήδη σήμερα ορισμένοι σύνθετοι και πλέον σημαντικοί μορφότυποι όπως το PDF και η HTML καθώς επίσης και παλαιότεροι μορφότυποι συμπεριλαμβανομένου του σκαναρισμένου εγγράφου, των σελίδων Unix, το TeX DVI, και το Apple II AppleWorks - επεξεργασία κειμένου. Η αρχιτεκτονική αυτή είναι σταθερή από τον καθορισμό της το 1997 και επεκτείνεται εύκολα στους επιπλέον νέους μορφότυπους εγγράφων αφού προσδιορίζεται από μια διασταύρωση της μορφής του δέντρου εγγράφου ως προς τη δομή των δεδομένων, η οποία συλλαμβάνει πλήρως τη σημασιολογία και το σχεδιάγραμμα και υποστηρίζει την πλήρη έκφραση των συχνά ιδιοσυγκρασιακών εννοιών ενός μορφότυπου, καθώς και τη συμπεριφορά, επιτρέπει το διαμοιρασμό της λειτουργικότητας μεταξύ των διαφόρων μορφότυπων και μειώνει έτσι την προσπάθεια εφαρμογής της. Επίσης μπορεί να εισαγάγει νέες λειτουργίες όπως οι υπερσύνδεσμοι και οι σημειώσεις στις παλαιότερες μορφές τύπων που δεν μπορούσαν να εκφράσουν τη λειτουργία αυτή και παρέχει μια ενιαία διεπαφή (API) σε όλους τους μορφότυπους.

Το πρότυπο των πολλαπλών χρήσεων είναι εντελώς αντίθετο με την εξομοίωση και τη μετατροπή που χρησιμοποιείτο ως σήμερα καθώς έχει υψηλού επιπέδου αρχιτεκτονική και εκτενή εφαρμογή .

1 Εισαγωγή

Από τα πολλά ζητήματα που αντιμετωπίζουμε στην ψηφιακή συντήρηση — όπως είναι η ανάγνωση των δεδομένων από τα παλαιά φυσικά μέσα (παραδείγματος χάριν το χαρτί) , ή συγκομιδή/συλλογή τους από τις ιστοσελίδες, η προέλευση των δεδομένων (από τι μέσο ή τι μορφότυπο δημιουργήθηκαν), τα μεταδεδομένα τους (και αν αυτά μπορούν να διατηρηθούν επίσης αναλλοίωτα), η διαχείριση των δεδομένων (από τα νέα μέσα), η μακροπρόθεσμη αποθήκευση τους, η διαθεσιμότητα τους, η πρόληψη καταστροφής τους, οι πολλαπλοί τύποι δεδομένων που υπάρχουν (όπως διάφορα επιστημονικά στοιχεία ή τα βίντεο), η προστασία της πνευματικής ιδιοκτησίας, και άλλα — αυτή η μελέτη εστιάζει στο πρόβλημα της πεπαλαιωμένης μορφής δεδομένων των ψηφιακών εγγράφων αφού εξακολουθεί να είναι ένα σημαντικό πρόβλημα. Όπως επισήμαναν οι κύριοι ομιλητές Rodig.P - Borghoff.U.M - Scheffczyk.J - Schmitz.L στη διάλεξη με θέμα την συντήρηση των ψηφιακών δημοσιεύσεων (πρακτικά του Συνεδρίου ACM για την εφαρμοσμένη μηχανική εγγράφων , το 2003) «... το πεπαλαιωμένο των μορφότυπων μέσων και των μορφότυπων δεδομένων είναι το πιο απαιτητικό πρόβλημα ενώ η συντήρηση των bitstreams μπορεί να ελεγχθεί/κυριαρχηθεί πλέον με τη χρησιμοποίηση των γνωστών τεχνικών. (δηλ. χρησιμοποίηση διαφόρων προσαρμογών μέσων , εξομοίωση των δεδομένων , μετανάστευση , κ.τ.λ.)»

Εστιάζοντας στους ψηφιακούς μορφότυπους εγγράφων, αξίζει να εξεταστεί τι αποτελεί την επιτυχή συντήρηση. Για τα έγγραφα αποτελούμενα από χαρτί, η επιτυχής συντήρηση του φυσικού υλικού υπονοεί ότι το περιεχόμενο θα μπορεί να εξεταστεί/διαβαστεί, έστω και μερικές φορές υπό περιορισμένη πρόσβαση. Οι ψηφιακές εργασίες διαφέρουν από τα έντυπα έγγραφα δεδομένου ότι η συντήρηση του ίδιου του υλικού τους, δηλ. των αρχείων δεδομένων τους, σχεδόν ολοκληρώνεται καθώς τα έγγραφα πρωτοσυλλέγονται (π.χ με τη διαδοχική αντιγραφή τους).

Παρόλα αυτά, η εξέταση των ψηφιακών εγγράφων είναι αρκετά πιο σύνθετη. Διότι πέρα από μερικούς μορφότυπους που βασίζονται σε κείμενο, όπως το κείμενο ASCII, όλοι οι άλλοι μορφότυποι εγγράφων είναι, εκτεταμένα εάν όχι εξ ολοκλήρου, αδύνατο να διαβαστούν χωρίς αποκωδικοποίηση από εξειδικευμένο λογισμικό. Επίσης, Τα ψηφιακά έγγραφα περιλαμβάνουν συχνά χρονοβασισμένο

περιεχόμενο (δηλαδή περιεχόμενο που έχει μια καθορισμένη χρονική διάρκεια, την οποία δεν γίνεται πολλές φορές να υπερβούμε) όπως οι ήχοι, το βίντεο, και τα κινούμενα γραφικά. Τα ψηφιακά έγγραφα περιλαμβάνουν συχνά και ενεργά στοιχεία όπως είναι οι διάφορες μορφές/φόρμες, τα γραφικά, και οι συνδέσεις. Στα πλαίσια κάποιων εκ των επιστημονικών δεδομένων και του λογισμικού, που παραλληλίζουν τα έγγραφα με τα ενσωματωμένα προγράμματα, ο Messerschmitt (σε διμηνιαία αναφορά που παρατίθεται η έκθεσή του, στο επιστημονικό περιοδικό ARL , σελίδα 229 , με τίτλο : Ευκαιρίες για τις ερευνητικές βιβλιοθήκες στο πρόγραμμα Κυβερνο-υποδομής NSF, το 2003) επισημαίνει ότι η διάκριση μεταξύ αυτών των δύο (δηλαδή μεταξύ της συντήρησης από την μια των επιστημονικών στοιχείων και από την άλλη της συντήρησης των στοιχείων του λογισμικού) «όλο και ταχύτερα γίνεται δυσδιάκριτη» και δηλώνει ότι «οι στόχοι συντήρησης των δεδομένων και οι στόχοι συντήρησης του λογισμικού δεν είναι χωριστοί/διαφορετικοί, αλλά πρέπει να υποτίθεται από την αρχή ότι είναι κατά ένα μεγάλο μέρος αδιάσπαστοι/κοινοί». Ο Marshall και ο Golonchinsky [διάλεξη με θέμα : Διάσωση του ιδιωτικού υπερκειμένου: Απαιτήσεις και πραγματικές διαστάσεις για τη συντήρηση - Πρακτικά του συνεδρίου ACM για το υπερκείμενο (2004) σελίδες 130–138] εξετάζουν τις επιπρόσθετες σταδιακές διαστάσεις/αποχρώσεις των λογοτεχνικών υπερκειμένων, που όπως αναφέρουν «προκύπτουν και από τον τρόπο εμφάνισης των εργασιών στην οθόνη μας , και από τη διαδραστική συμπεριφορά τους και από τους τρόπους που καταγράφεται η αλληλεπίδραση των αναγνωστών σε σχέση με την εργασία που κάνουν».

Παρόμοια επιχειρήματα θα μπορούσαν να προβληθούν και από άλλους ερευνητές για πολλά ερευνητικά συστήματα γενικώς και ακόμη για οποιοδήποτε άλλο σύστημα εγγράφων που χαρακτηρίζεται από διαφορετικές ιδιοσυγκρασιακές έννοιες. Επίσης, η συντήρηση/διατήρηση της δυνατότητας της εξέτασης κάποιων βασικών στοιχείων του λογισμικού ή κάποιων κύριων στοιχείων μερικών εγγράφων είναι ιδιαίτερα προβληματική όταν το λογισμικό αυτό είναι ιδιόκτητο (άρα δεν έχουμε πάντα την επιθυμητή πρόσβαση σε αυτό), εάν το λογισμικό λειτουργεί μόνο με ξεπερασμένο/πεπαλαιωμένο υλικό και αν οι μορφότυποι των δεδομένων δεν είναι δημόσιοι μορφότυποι (π.χ PDF).

Και η εξέταση ενός εγγράφου (το να το δεις και να μπορείς να το διαβάσεις) δεν είναι αρκετή. Είναι αναμενόμενες περισσότερες επιπλέον λειτουργίες από τα ψηφιακά έγγραφα από τα τυπωμένα σε χαρτί έγγραφα . Λόγου χάρη, οι χρήστες

στα ψηφιακά έγγραφα προσδοκούν να μπορούν να αντιγράψουν και να επικολλήσουν (σε άλλη θέση) το κείμενο, τις εικόνες, τα βίντεο, και οποιουδήποτε άλλους τύπους περιεχομένου . Τα ιδρύματα, οι διάφοροι ιστοχώροι, μέχρι και τα άτομα , όλοι προσδοκούν να δύνανται να ψάξουν στο περιεχόμενο των εγγράφων [αυτή η λειτουργία θεωρείται τόσο θεμελιώδης που πλέον χτίζεται (δηλαδή ενσωματώνεται) στα λειτουργικά συστήματα επόμενης-γενιάς] . Για ορισμένους χρήστες (όπως είναι π.χ τα άτομα με κινητικά προβλήματα ή με προβλήματα όρασης, ή ακόμη και άτομα που ο χρόνος τους λόγω διαφόρων άλλων υποχρεώσεων είναι περιορισμένος) οι λογογράφοι/κειμενογράφοι και αυτόματη γραφή Braille είναι πολύ ουσιαστικές λειτουργίες . Οι επιχειρήσεις και οι ερευνητές θέλουν να έχουν τη δυνατότητα να μπορούν να εκτελέσουν εξόρυξη των κειμένων με αυτόματη γλωσσική μετάφραση - ώστε να γίνονται άμεσα κατανοητά τα δεδομένα και να μην υπάρχει καμία χρονική καθυστέρηση ή επιπλέον οικονομικό κόστος. Επίσης, οι ερευνητές θέλουν (να συγκεντρώνουν πληροφορίες και) να συμπεραίνουν έως και τη σημασιολογική δομή των εγγράφων, χρησιμοποιώντας όλες τις πληροφορίες που περιέχει το έγγραφο. Κι όταν λέμε όλες τις πληροφορίες, εννοούνται πραγματικά όλες, δηλαδή, από τις συντεταγμένες του σχεδιαγράμματος ως τα τελικά σχέδια των φύλλων (αν υπάρχουν) έτσι ώστε να απλοποιηθεί η ρητή σημασιολογική δομή τους (αν υπάρχει). Ακόμα και οι απλοί άνθρωποι/χρήστες θέλουν να μπορούν να μετατρέπουν τα έγγραφα που τους ενδιαφέρουν στο μορφότυπο που τους εξυπηρετεί κάθε δεδομένη στιγμή της ημέρας, (όπως γίνεται π.χ με τις φορητές συσκευές με τις μικρές οθόνες – π.χ PDA's). Αυτοί οι χρήστες θέλουν να μπορούν να προσθέσουν υπερσυνδέσμους στα δεδομένα και να σχολιάζουν/σημειώνουν πάνω σε αυτά, ακόμα κι αν ο μορφότυπος του εγγράφου δεν υποστηρίζει αυτές τις αυτές τις λειτουργίες.

Μπορούμε να προσδοκούμε ότι το μέλλον θα είναι όλο και περισσότερο απαιτητικό καθώς νέες εφαρμογές εφευρίσκονται σχεδόν καθημερινά, οι οποίες στηρίζονται επάνω σε ενδεχομένως οποιουδήποτε περιεχομένου έγγραφα .

2 Σχετικές εργασίες

2.1 η Συντήρηση του υλικού

Το λογισμικό που χρησιμοποιείται για την δημιουργία και την εξέταση των εγγράφων, όπως οποιοδήποτε λογισμικό γενικότερα, πολύ συχνά απαιτεί συγκεκριμένο ενισχυτικό λογισμικό (τις γνωστές αναβαθμίσεις λογισμικού) και άμεσα ή έμμεσα (έμμεσα ίσως μέσω του λειτουργικού συστήματος) να απαιτεί συγκεκριμένο υλικό. Το πρόβλημα είναι ότι καθώς το υλικό αυτό χαλάει/διαλύεται (είτε με την πάροδο του χρόνου είτε λόγω της χρήσης που υφίσταται) το υλικό της νέας γενιάς που χρησιμοποιείται τη δεδομένη εποχή, μπορεί να μην είναι συμβατό με το λογισμικό που προϋπάρχει .

Η Συντήρηση του υλικού μπορεί και συντηρεί ταυτόχρονα την αναγνωσιμότητα του αρχικού ψηφιακού εγγράφου, διατηρώντας το αρχικό υλικό και το δεδομένο λογισμικό για αόριστο χρόνο. Όμως, ένα τέτοιο μουσειακό/πετρελαιωμένο υλικό προορίζεται τελικά να αποτύχει, καθώς αφού το υλικό κάποια στιγμή θα χαλάσει εάν δεν έχουμε κάποιο άλλο παρόμοιο μηχάνημα/σύστημα που να μπορούμε να το σπάσουμε/διαλύσουμε για να βρούμε τα επιμέρους τμήματα/μέρη που χρειάζεται να αντικαταστήσουμε στο πετρελαιωμένο υλικό, δεν θα μπορέσουμε να το συντηρήσουμε. Εάν λάβουμε υπόψη μάλιστα το γεγονός ότι τα μέρη/τμήματα αυτά που θα είναι απαραίτητο να γίνει αντικατάσταση τους στο παλαιό λογισμικό είναι πάρα πολύ εξειδικευμένα (συνήθως είναι λίγα και προφανώς αφορούν μεμονωμένα/συγκεκριμένα μέρη του συστήματος το κάθε ένα από αυτά που θα απαιτηθεί) ώστε να προβούν οι εταιρείες παραγωγής τους στη διάθεση πόρων για την επανάληψη της κατασκευής τους επικερδώς , συνειδητοποιούμε τη βέβαιη αποτυχία του .

2.2 Εξομοίωση

Το απαραίτητο/αρχικό υλικό όμως μπορεί να προσαρτηθεί/αντιγραφεί στο λογισμικό των τωρινών (ισχυρότερων) υπολογιστών, και επομένως οι εξομοιωτές μπορούν να αναπαραγάγουν την ακριβείς εμφάνιση και τη συμπεριφορά ενός εγγράφου. Βέβαια, ας μην ξεχνάμε ότι απαιτεί αρκετή εργασία από τους εμπειρογνώμονες για να καταφέρουν να μιμηθούν έναν υπολογιστή, ειδικά έναν σύγχρονο υπολογιστή, αλλά το καλό είναι ότι υπάρχουν στην αγορά διαθέσιμες πολλές εφαρμογές για τέτοιους εξομοιωτές και πωλούνται από διάφορες επιχειρήσεις. Έτσι, όταν ο τρέχων δεδομένος υπολογιστής γίνεται ξεπερασμένος, ένας νέος εξομοιωτής θα μπορεί να τρέξει αντί γι'αυτόν τον εξομοιωτή της προηγούμενης γενεάς (ο οποίος ήταν συμβατός με το δεδομένο υπολογιστή), και τις λοιπές εφαρμογές, δημιουργώντας κατά αυτόν τον τρόπο έναν όλο και συνεχώς αναπτυσσόμενο σωρό εξομοιωτών, ο οποίος μπορεί να είναι ή να μην είναι εν τέλει βιώσιμος. Και αυτό (δηλαδή το να μην είναι τελικά βιώσιμος) θα συμβεί για τον εξής λόγο: διότι σε πολλές περιπτώσεις που αντιμετωπίστηκαν έως τώρα, το περιεχόμενο του εγγράφου παρέμεινε παγιδευμένο μέσα στον εξομοιωτή και κάπου εκεί μέσα στη «μνήμη-σούπα» του εξομοιωτή (την αναφέρουμε ως σούπα αφού στη μνήμη του εξομοιωτή θα υπάρχουν ανακατεμένοι πολλοί μορφότυποι και στοιχεία δεδομένων) βρίσκονται και οι δομές των δεδομένων του προγράμματος που αντιπροσωπεύουν ουσιαστικά "το έγγραφο". Πέρα από αυτά, το γεγονός να γίνει εφικτό να βρεθεί το έγγραφο (μέσα στη μνήμη του εξομοιωτή) και να γίνει η εξαγωγή του σε άλλο νέο σύστημα, παραμένει τουλάχιστον τόσο δύσκολο όσο το να προσπαθήσουμε να αποσαφηνίσουμε το αρχικό bitstream των δεδομένων των αρχείων του εγγράφου. Θα ήταν πολύ επιθυμητό να καταφέρουμε να προσθέσουμε το περιεχόμενο του εγγράφου σε μια μηχανή αναζήτησης ή να στείλουμε το έγγραφο σε άλλους να το διαβάσουν χωρίς το λειτουργικό κόστος του υπερφορτωμένου εξομοιωτή, αλλά δυστυχώς δεν μπορούμε.

2.3 Μετατροπή/μετανάστευση

Η μετατροπή, αποκαλούμενη επίσης και μετανάστευση, έχει την ικανότητα να δέχεται το υλικό με έναν παλαιότερο σχήμα/μορφότυπο και να το κωδικοποιεί ξανά εκ νέου σε ένα νεότερο μορφότυπο. Αυτή η διαδικασία μπορεί να έχει κάποια επιτυχία για τα απλά δεδομένα (λόγου χάρη, ίσως οι πολλαπλοί μορφότυποι των εικόνων να μπορούν όλοι να αντιπροσωπευθούν σε ένα δυσδιάστατο πλέγμα τιμών χρώματος). Αλλά τα ψηφιακά έγγραφα είναι πιο σύνθετα και γενικά σημασιολογικά είναι ασυμβίβαστα μεταξύ τους, και οι μετατροπές από τον ένα μορφότυπο σε άλλο μορφότυπο σχεδόν πάντα χάνουν σε πληροφορίες, για το θεμελιώδη λόγο ότι μερικές έννοιες που εκφράζουν έναν μορφότυπο εγγράφου δεν μπορούν πάντα να εκφραστούν σε άλλο (μορφότυπο).

Δεδομένου μάλιστα του γεγονότος ότι ένας μορφότυπος εγγράφου εξελίσσεται κάθε λίγα χρόνια με τις νέες ενημερώσεις του αντίστοιχου λογισμικού, το λογισμικό μπορεί συνήθως να διαβάσει τις 2-4 τελευταίες εκδόσεις του μορφότυπου του ,αλλά τα έγγραφα που είναι παλαιότερα από μερικά μόνο έτη μπορεί να γίνουν δυσανάγνωστα . Κατά συνέπεια εικάζουμε ότι η διαδικασία της μετατροπής θα απαιτεί πάντα τη σταθερή προσοχή, θα απαιτεί σταθερή μετανάστευση. Αυτή η αλυσίδα μετατροπής από μορφότυπο σε μορφότυπο μπορεί να χάσει πληροφορίες σε κάθε βήμα, υποβιβάζοντας αδυσώπητα την ποιότητα. Επιπλέον, ενώ η απώλεια δεδομένων είναι σχεδόν εγγυημένη για τις μετατροπές μεταξύ των μορφότυπων εγγράφων, είναι ιδιαίτερα πιθανή ακόμη και για τις αναβαθμίσεις λογισμικού στην ίδια εφαρμογή.

Σήμερα, παρόλο που η εξομοίωση και η μετατροπή υποφέρουν σαν λειτουργίες καθώς υφίστανται στα γνωστά προβλήματα (που προαναφέρθηκαν), θεωρούνται συχνά ως οι μόνοι τρόποι ικανής συντήρησης των ψηφιακών εγγράφων. Τα βρετανικά εθνικά αρχεία προσπαθούν να μετριάσουν τη ζημία που γίνεται (την απώλεια δηλαδή των πληροφοριών και την υποβάθμιση της ποιότητας κατά τις διαδικασίες μετατροπής ή εξομοίωσης) με την ανάπτυξη μιας βάσης δεδομένων των μορφότυπων αρχείων, η αποκαλούμενη PRONOM, η οποία βάση δεδομένων «...επιτρέπει την αυτόματη παραγωγή των απαιτούμενων διαβάσεων μετανάστευσης/μετατροπής, με το να προσδιορίζει την κάθε πιθανή διαδρομή μετατροπής που θα εμφανιστεί μεταξύ ενός πηγαίου/αρχικού μορφότυπου και ενός μορφότυπου στόχου, παρέχοντας όσο το δυνατόν

περισσότερες πληροφορίες για το πώς το κάθε στάδιο μετατροπής θα έχει επιπτώσεις στο αρχικό περιεχόμενο του εγγράφου...» (Πρακτικά του Συνεδρίου για την Απευθείας Σύνδεση –online– με την Πληροφορία (2003), διάλεξη : Brown.A με θέμα : η Συντήρηση της ψηφιακής κληρονομιάς : η Οικοδόμηση ενός ψηφιακού αρχείου για τα αρχεία βρετανικής κυβέρνησης).

[Η βάση αυτή πάραυτα βρίσκεται ακόμη στο στάδιο της ανάπτυξης (δεν έχει αναπτυχθεί πλήρως και ολοκληρωμένα – οι ερευνητές αναφέρουν ότι κάτι τέτοιο ίσως να μην είναι ποτέ εφικτό μιας και δημιουργούνται ταχύτατα και συνεχώς νέοι μορφότυποι) όμως τα αναμενόμενα αποτελέσματα της εφαρμογής της θεωρούνται ένα πολύ σημαντικό βήμα και για την εξέλιξη της συντήρησης των ψηφιακών εγγράφων αλλά και για την ανάπτυξη παρόμοιων εφαρμογών στο εγγύς μέλλον.]

Εντούτοις, ακόμα κι αν η ζημία που γίνεται σε κάθε βήμα είναι περιορισμένη, όταν αυτή πολλαπλασιάζεται με τις δεκάδες ή τις εκατοντάδες μετατροπές που λαμβάνουν χώρα κατά τη διάρκεια των ετών, και απ΄τη στιγμή που μια τέτοια χρονική επέκταση είναι τελικά το θέμα (ζητούμενο) της συντήρησης εξ αρχής , η απώλεια των δεδομένων είναι και σημαντική και σίγουρη.

Το πρόγραμμα CAMiLEON [Creative Archiving at Michigan and Leeds Emulating the Old on the New : Το πρόγραμμα CAMiLEON αναπτύσσει και αξιολογεί μια σειρά στρατηγικών τεχνικών για τη μακροπρόθεσμη συντήρηση των ψηφιακών υλικών. (οι μελέτες αξιολόγησης των χρηστών και μια ανάλυση των δαπανών συντήρησης δίνουν τις απαντήσεις όσον αφορά το «πότε» και «που» θα χρησιμοποιηθούν αυτές οι στρατηγικές) Το πρόγραμμα είναι κοινή εργασία μεταξύ των πανεπιστημίων του Μίτσιγκαν (ΗΠΑ) και του Λίντς (Ηνωμένο Βασίλειο) και χρηματοδοτείται από το JISC και το NSF. Το πρόγραμμα αυτό έχει τρεις κύριους στόχους: 1. να ερευνήσει τις επιλογές για τη μακροπρόθεσμη διατήρηση της αρχικής λειτουργικότητας, της εικόνας καθώς και αίσθησης των ψηφιακών αντικειμένων 2. να ερευνήσει την τεχνολογική εξομοίωση ως μακροπρόθεσμη στρατηγική για τη μακροπρόθεσμη συντήρηση και την πρόσβαση στα ψηφιακά αντικείμενα, 3. να εξετάσει το που και το πώς αρμόζει η εξομοίωση σε ένα σύνολο από στρατηγικές ψηφιακής συντήρησης . Κάποια από τα κύρια προϊόντα του προγράμματος θα είναι : ένα σύνολο εργαλείων συντήρησης που θα είναι διαθέσιμο για χρήση και για δοκιμή στις βιβλιοθήκες, συγκρίσεις δαπανών των διαφορετικών επιπέδων εξομοίωσης δεδομένου ότι ίσως να χρησιμοποιηθεί ως βασική στρατηγική για τη μακροπρόθεσμη συντήρηση, οδηγίες για τη χρήση των

διαφορετικών στρατηγικών τεχνικών (μετανάστευση και εξομοίωση) για τη συντήρηση των ψηφιακών συλλογών, ορισμοί των ιδιοτήτων των διαφορετικών τύπων ψηφιακών αντικειμένων που πρέπει να συντηρηθούν για να ικανοποιήσουν τις ανάγκες και τις απαιτήσεις των χρηστών, στρατηγικές για τα ψηφιακά αντικείμενα για τα οποία δεν υπάρχει καμία γνωστή μέθοδος μακροπρόθεσμης συντήρησης και πρόσβασης] εξετάζει το συσσωρευτικό πρόβλημα της απώλειας των δεδομένων με το να τα μετατρέπει πάντα από το αρχικό bytestream . Τα έγγραφα διαβάζονται σε ένα ενδιάμεσο μορφότυπο και διάφοροι άλλοι νέοι μορφότυποι μετατροπών μπορούν να αναπτυχθούν εάν απαιτηθεί. Η αρχιτεκτονική αυτή καταδείχθηκε (παρουσιάστηκε) σε μια επιλογή διανυσματικών μορφότυπων σε γραφική παράσταση . Αυτό είναι μεν ελπιδοφόρο, αλλά αντιμετωπίζει πρόσθετα ζητήματα/προβλήματα όταν εφαρμόζεται στα πιο σύνθετα έγγραφα. Διότι, ακόμη και μεταξύ των διανυσματικών μορφότυπων της γραφικής παράστασης, τα σημασιολογικά χάσματα προαπειτούσαν τα στοιχεία να υποβιβασθούν, και μπορούμε να περιμένουμε και περισσότερα προβλήματα πέρα από αυτό (δηλαδή, ακόμη και πλήρης απώλεια δεδομένων σε κάποιες περιπτώσεις) ειδικά με τα σύνθετα πρότυπα εγγράφων, στα οποία μπορεί να είναι ή μπορεί και να μην είναι η αρχιτεκτονική αυτή ένας αποδεκτός συμβιβασμός. Επίσης, δεν εξετάζει/καταδεικνύει τη συμπεριφορά των εγγράφων(όπως κάνει π.χ ο χειρισμός με JavaScript ενός HTML DOM) . Το ενδιάμεσο σχήμα που προτείνει φαίνεται να είναι ένα συνολικά ενωτικό σχήμα των εννοιών από όλους τους υποστηριζόμενους μορφότυπους και από πρακτική σκοπιά θα γινόταν πιθανώς υπερβολικά μεγάλο και αδέξιο καθώς εκατοντάδες ή χιλιάδες μορφότυποι εγγράφων θα υιοθετούνταν, πολλοί από αυτούς με ιδιοσυγκρασιακές έννοιες και κυρίως οι περισσότεροι με αναρίθμητες μικρές αλλά ενδεχομένως σημαντικές παραλλαγές στις κοινές τους δομές .

2.4 το Καθολικό(Παγκόσμιο) σχήμα

Μερικά συστήματα μετατρέπουν όλες τις πηγές σε ένα ενιαίο καθολικό σχήμα/μορφότυπο , το οποίο χρησιμοποιούν για όλες τις περαιτέρω διαδικασίες. Η γλώσσα XML φαίνεται να είναι ένας ελκυστικός υποψήφιος δεδομένου ότι συλλαμβάνει και τη σημασιολογία και τη δομή, είναι επεκτάσιμη και είναι εύκολο στο να αναλυθεί. Το Εικονικό έγγραφο [αυτό το πρότυπο εγγράφου επιτρέπει τη μετατροπή ενός απλού –συνήθως PDF– εγγράφου σε ένα εικονικό έγγραφο το οποίο μπορεί να δημοσιευτεί απευθείας –online– στον Παγκόσμιο Ιστό, να εκτυπωθεί ή να αποθηκευτεί σε κάποιο φυσικό μέσο, κ.τ.λ. Το βασικό χαρακτηριστικό του όμως είναι η διαδραστικότητα που μπορεί να του προσδώσει κάποιες ξεχωριστές ιδιότητες που δεν ίσχυαν για τους απλούς μορφότυπους εγγράφων (αλλά τις συναντούσαμε συχνά σε ιστοσελίδες) όπως είναι η εισαγωγή (μέσο κάποιων εργαλείων του προτύπου) αρχείων βίντεο και μουσικής, η ύπαρξη καλαθίου αγορών (για online περιοδικά ή καταλόγους π.χ), στατιστικές μετρήσεις, υπηρεσία ηλεκτρονικού ταχυδρομείου που λειτουργεί μέσα στο έγγραφο, και όλα αυτά σε πραγματικό χρόνο] και το UrLib [προσωπικό ψηφιακό σύστημα βιβλιοθηκών : παρέχει μια ασφαλή μακροπρόθεσμη αποθήκευση και ένα σύστημα ανάκτησης για μια μεγάλη γκάμα (προσωπικών) εγγράφων όπως είναι τα έγγραφα, οι φωτογραφίες, τα βιβλία, το ηλεκτρονικό ταχυδρομείο κ.α. Το σύστημα δημιουργεί ένα ολοκληρωμένου-κειμένου (full text) αποθετήριο που γίνεται προσβάσιμο διαμέσου ενός ενεργού πράκτορα που λειτουργεί μέσω μιας διεπαφής (interface) του Ιστού. Είναι κατάλληλο για συλλογές που περιλαμβάνουν χιλιάδες εγγράφων και επιτρέπει τόσο την ευκολία στην είσοδο εγγράφων όσο και την πρόσβαση, καθώς επίσης και υψηλά επίπεδα ασφάλειας και μυστικότητας για το χρήστη. Είναι χρήσιμο ως πλατφόρμα για την περαιτέρω έρευνα σε ψηφιακές βιβλιοθήκες και σε υπολογιστές όπου εκπληρώνεται αυξημένη ανάγνωση ψηφιακών εγγράφων.] (κανένα από τον οποία δεν αξιώνεται/ισχυρίζεται να είναι μια βάση για την ψηφιακή συντήρηση) λύνουν το πρόβλημα του πολλαπλού μορφότυπου με το να συλλαμβάνουν την εικόνα και τις αντιπροσωπεύσεις σε κείμενο όλων των εγγράφων, τηρώντας μια "προβολή" στην θέα (στο να βλέπεις) και άλλη στην αναζήτηση.

Τα διασημότερα παραδείγματα των καθολικών μορφότυπων είναι το PostScript και το PDF, τα οποία καυχώνται το μοναδικό πλεονέκτημα ότι μπορούν

ήδη να συλλάβουν οποιοδήποτε έγγραφο που μπορεί να τυπωθεί (σε χαρτί συνήθως – που είναι βασικά όλοι οι μορφότυποι με στατικό περιεχόμενο) και ολοένα και περισσότερες εφαρμογές πλέον παράγουν αρχεία PDF και άμεσα και σε ένα υψηλό σημασιολογικό επίπεδο σε σχέση με εκείνα που στέλνονται μέσω (του οδηγού) των εκτυπωτών. Σε ένα ενιαίο σχήμα σαν αυτό που αναφέρεται, το πρότυπο PDF υποστηρίζει υψηλής πιστότητας εικόνα (αυτό που βλέπει κανείς) καθώς επίσης και κείμενο-βασισμένες διαδικασίες όπως η έρευνα . Ακόμη, το πρότυπο αρχείων PDF μπορεί να ενσωματώσει το bitstream του αρχικού εγγράφου για μελλοντική επεξεργασία ή πιο απαιτητική συντήρηση. Η εταιρεία Adobe προβάλλει/λανσάρει το PDF για την αρχειοθέτηση, επισημαίνοντας ότι το PDF είναι δημόσια διαθέσιμο πρότυπο (αλλά όχι ανοιχτό, δηλαδή λογισμικό ανοιχτού κώδικα) και υποστηρίζει τα αρχεία μεταδεδομένων της γλώσσας XML, μεταξύ των άλλων χαρακτηριστικών του. Ένα μετά-πρότυπο του PDF για την αρχειοθέτηση ονομάζεται PDF/A και το οποίο μπορεί να προσδιορίζει/αναγνωρίζει " το σύνολο των PDF συστατικών/στοιχείων που μπορούν να χρησιμοποιηθούν από μια εφαρμογή και τους περιορισμούς στη μορφή της χρήσης τους " (όπως για παράδειγμα είναι η μη αποδοχή-δεν το επιτρέπει-του κατοχυρωμένου με δίπλωμα ευρεσιτεχνίας -πατέντα- φίλτρου συμπίεσης LZW μιας και υφίσταται η προαπαιτήση ότι όλες οι πηγές θα ενσωματώνονται) .

Κατά κάποιο τρόπο το καθολικό σχήμα είναι αιώνιο, και ίσως γίνει τόσο σημαντικό που η ίδια η κοινωνία θα το εξασφαλίζει . Εντούτοις, η προσέγγιση αυτή έχει και τους περιορισμούς της. Όπως είναι λογικό , δεν είναι πρακτικό να συλληφθούν εντελώς όλες οι πτυχές όλων των μορφότυπων εγγράφων μέσα σε ένα ενιαίο καθορισμένο μορφότυπο . Διότι ο μορφότυπος αυτός θα ήταν αδέξιος, εχθρικός στην πλήρη εφαρμογή και θα έπρεπε να ενημερώνεται συνεχώς καθώς άλλοι νέοι μορφότυποι θα εισάγονται/δημιουργούνται . Οι αποκαλούμενοι λοιπόν καθολικοί μορφότυποι πρέπει λόγω πρακτικής ανάγκης να επιλέγουν ορισμένα μόνο χαρακτηριστικά (από τους μορφότυπους που θα συνδυάζουν) και να αφήνουν κάποια άλλα (να τα προσπερνούν) και έτσι λοιπόν θα υπάρχει ένα βήμα μετατροπής και μια αντίστοιχη απώλεια δεδομένων στη χρήση τους.

2.5 Καθολικός(Παγκόσμιος) εικονικός υπολογιστής

Ο Raymond Lorie προτείνει τους διερμηνείς γραπτών δεδομένων που « θα μπορούν να εξαγάγουν τα στοιχεία από το bitstream και να τα επιστρέψουν στον επισκέπτη με έναν κατανοητό τρόπο, έτσι ώστε να μπορούν να μεταφερθούν και σε ένα νέο σύστημα » [Lorie Raymond : Μακροπρόθεσμη συντήρηση των ψηφιακών πληροφοριών, Πρακτικά του πρώτου Συνεργατικού/κοινού Συνεδρίου μεταξύ ACM / IEEE-CS για τις Ψηφιακές Βιβλιοθήκες (2001) σελίδες 346–352]. Τα προγράμματα γράφονται σε συνδυασμό με έναν καθολικό εικονικό υπολογιστή (UVC=Universal Virtual Computer) έτσι ώστε στο μέλλον, όλα αυτά που θα απαιτούνται να είναι μια εφαρμογή του UVC στον υπολογιστή της ημέρας (τον απλό υπολογιστή μας δηλαδή) ώστε να τρέξουν οι διερμηνείς και να διαβάσουν έτσι τα δεδομένα.

[Η έννοια του καθολικού εικονικού υπολογιστή αναπτύχθηκε από τον Raymond Lorie. Είναι ένα σύστημα που χρησιμοποιεί τα στοιχεία και της μετανάστευσης και της εξομοίωσης και επιτρέπει στα ψηφιακά αντικείμενα να ανασυγκροτηθούν με την αρχική μορφή τους. Η έννοια του UVC αποτελείται από : τον ίδιο τον UVC , (που είναι ένας λογικός μορφότυπος δεδομένων με περιγραφές των τύπων) το πρόγραμμα UVC (αποκωδικοποίησης μορφότυπων) και τον λογικό αναγνώστη δεδομένων. Ο UVC έχει ήδη πραγματοποιηθεί και έχει αποδειχθεί ικανός να εργάζεται για τους τύπους εικόνας χρησιμοποιώντας τα πρότυπα JPEG και GIF . Εντούτοις, παρατηρείται ότι η απόδοση του παρόντος UVC δύναται ακόμα να βελτιωθεί περισσότερο. Το εργαλείο επίδειξης του (demonstration tool) UVC είναι ελεύθερα διαθέσιμο στην ιστοσελίδα Alphaworks για να επιτρέψει στο ευρύ κοινό να πειραματιστεί/ανακαλύψει αυτή τη προσέγγιση της μακροπρόθεσμης πρόσβασης].

Ο UVC είναι εξαιρετικά προσεκτικός για εκείνα που θεωρούνται να είναι σίγουρα για το μέλλον και απαιτεί ευτυχώς απλώς λίγα περισσότερα στοιχεία από το αντίστοιχο ενός απλού μικροεπεξεργαστή και μιας μνήμης. Λαμβάνοντας υπόψη αυτήν την προσέγγιση από την πρακτική σκοπιά των μηχανικών του λογισμικού οι οποίοι είναι υπεύθυνοι για την οικοδόμηση/δημιουργία ενός συστήματος που θα αγκαλιάζει/ενσωματώνει εκατοντάδες μορφότυπους εγγράφων οι οποίοι είναι μερικές φορές πολύ πολύπλοκοι, αυτό δεν είναι αρκετό. Στην πράξη, οι μηχανικοί λογισμικού χρειάζονται μια αρχιτεκτονική που να

περιγράφει τη μεγάλης κλίμακας οργάνωση του λογισμικού που θα χτιστεί και που θα εξηγεί λεπτομερώς τις αλληλεπιδράσεις μεταξύ των πολλών συστατικών/στοιχείων. Για τη συντήρηση των ψηφιακών εγγράφων, αυτή η αρχιτεκτονική θα πρέπει να αγκαλιάσει/ενσωματώσει ορισμένες συγκεκριμένες (από το πεδίο/περιοχή τους) έννοιες όπως "το έγγραφο", "τα μεταδεδομένα", "το κείμενο", "τη συμπεριφορά" και τέλος "τη δομή".

Στην πράξη, οι μηχανικοί λογισμικού χρειάζονται/απαιτούν μια γλώσσα υψηλού επιπέδου, όπως η γλώσσα Java και επιπλέον βιβλιοθήκες προκαθορισμένων λειτουργιών (οι οποίες όλες να μπορούν να συνάγονται / συναθροιστούν στον UVC). Ένα επίπεδο επάνω από τον UVC θα πρέπει να διασυνδεθεί με το υλικό, όπως οι οθόνες, τα πληκτρολόγια, και τα ποντίκια. Ο UVC του Logie είναι μια σταθερή αρχή και τώρα ήρθε ο καιρός για την ακόμα υψηλότερου επιπέδου αρχιτεκτονική και την εφαρμογή της.

3 Τα οφέλη της Πολυχρηστικής Αρχιτεκτονικής για τη Συντήρηση

Εξετάζουμε λοιπόν το πρότυπο εγγράφων Multivalent¹ για να δούμε πώς οι αρχιτεκτονικές ιδιότητές του υποστηρίζουν την ψηφιακή συντήρηση εγγράφων. Ο σκοπός εδώ δεν είναι μια παρουσίαση της αρχιτεκτονικής αυτής καθ' εαυτής, αλλά να διευκρινίσουμε/αποσαφηνίσουμε το πώς οι σημαντικές πτυχές του προβλήματος της ψηφιακής συντήρησης λύνονται από ορισμένες πτυχές της αρχιτεκτονικής, μερικές φορές μάλιστα αποκλειστικά/μοναχά έτσι .

Η αρχιτεκτονική αυτή είναι ισχυρή και ευπροσάρμοστη όπως μπορεί να εκτιμηθεί από την ακόλουθη περιγραφή μιας προηγούμενης εφαρμογής της αρχιτεκτονικής αυτής σε έναν φυλλομετρητή του Ιστού.

[Ο πολυχρηστικός φυλλομετρητής στηρίζεται σε μια αρχιτεκτονική που διαχωρίζει τη λειτουργικότητα του εγγράφου από το σταθερό μορφότυπο των εγγράφων. Σχεδόν όλη αυτή η λειτουργικότητα διατίθεται μέσω κάποιων σχετικά μικρών ενοτήτων/καταμητών του κώδικα του προγράμματος, που αποκαλούνται συμπεριφορές τις οποίες οι προγραμματιστές μπορούν να τις επαναγράψουν για να επεκτείνουν εάν επιθυμούν μέχρι και το κεντρικό/βασικό σύστημα. Οι συμπεριφορές μπορούν να είναι ιδιαίτερα σημαντικές και ισχυρές αλλά άλλο τόσο καινοτόμες ή ασυνήθιστες. Έχει ήδη σχεδιαστεί και εφαρμοστεί αυτή η αρχιτεκτονική (σε αυτήν βασίζεται εξάλλου το πρότυπο του πολυχρηστικού φυλλομετρητή) και έχουν εφαρμοστεί σε αυτήν συμπεριφορές που υποστηρίζουν όχι μόνο πλήρως την λειτουργικότητα (των εγγράφων) αλλά και ακόμη περισσότερα!] Ο πολυχρηστικός φυλλομετρητής καταμετράει πολλούς τοπικούς μορφότυπους εγγράφων (όπως είναι οι εξής : PDF, HTML, σκαναρισμένα έγγραφα, τις σελίδες Unix, το TeX DVI και άλλα) και υποστηρίζει τον επί τόπου σχολιασμό (επισημάνσεις, σημειώσεις, εκτελέσιμη αντιγραφή - σήμανση του συντάκτη, υποσημειώσεις) σε όλους τους μορφότυπους. Οι σημειώσεις αυτές μπορούν να συνδεθούν με οποιοδήποτε σημείο ενός εγγράφου (είτε πρόκειται για κείμενο είτε για εικόνα), επίσης μπορούν να εφαρμοστούν για τα έγγραφα που είναι μόνο για ανάγνωση (όπως π.χ η αρχική/κεντρική σελίδα της εφημερίδας New York Times), διασφαλίζονται με τις σταθερές τοποθεσίες έτσι ώστε να μπορούν να ξανά-προσαρτηθούν σωστά ακόμα κι αν το πηγαίο έγγραφο έχει επεξεργαστεί εκτενώς και εκμεταλλεύονται τους σταθερούς υπερσυνδέσμους

υπερκειμένου για να βρουν ένα έγγραφο εάν μετακινείται κάπου αλλού στο διαδίκτυο.

Καθώς η αρχιτεκτονική έχει εφαρμοστεί, το σύστημα καταμετράει ήδη συνολικά άνω των 100.000 γραμμών και πάνω από 4 εκατομμύριων χαρακτήρων του πηγαίου κώδικα. Οι προαναφερθέντες κατατμητές εγγράφων και ο φυλλομετρητής είναι ελεύθερα προσβάσιμοι στο διαδίκτυο [στην ιστοσελίδα του μοντέλου Multivalent : <http://multivalent.sourceforge.net> , κατηγορία Downloads]. Κατά το προηγούμενο έτος, η εφαρμογή αυτή χρησιμοποιήθηκε για τη συντήρηση, πρώτη φορά στη πόλη Σαν Ντιέγκο της Αμερικής στο ερευνητικό κέντρο υπερυπολογιστών και συγκεκριμένα στο πρόγραμμα δοκιμών επίμονων αρχείων [San Diego Supercomputer Center. Persistent Archive Testbed (PAT) <http://www.sdsc.edu/PAT/>].

Η αρχιτεκτονική αποδεικνύεται κατά τη διάρκεια του χρόνου. Από το 1997, καθώς η ενιαία διεπαφή API έχει εξελιχθεί και η εφαρμογή έχει προωθηθεί, η πολυχρηστική αρχιτεκτονική έχει παραμείνει σταθερή. (και ενώ προηγείται χρονικώς/προχρονολογείται από τη UVM του Lorie, προξένησε τον Lorie να επισημάνει έμμεσα την καταλληλότητά του για την ψηφιακή συντήρηση) .

Η αρχιτεκτονική έχει πολλές έννοιες ενδοασφάλισης και μπορεί να είναι καθοδηγητική ιδιαίτερα αρχικά καθώς εν συντομία εξετάζει/λαμβάνει υπόψη το σύνολο. Νέοι μορφότυποι εγγράφων υποστηρίζονται από τους προσαρμογείς μέσων, οι οποίοι είναι τμήματα του κώδικα που μεταφράζουν τους υπαρκτούς/σταθερούς μορφότυπους εγγράφων σε δομές δεδομένων τρέχοντα χρόνου . Η αρχική δομή των δεδομένων είναι το δέντρο του εγγράφου, που αντιπροσωπεύει ολόκληρο το περιεχόμενο ενός εγγράφου (με την κύλιση-της σελίδας- ή με μια σελίδα τη φορά), συμπεριλαμβάνει τα πάντα, από το κείμενο και τις εικόνες μέχρι τα χειρόγραφα, και μέχρι τη σημασιολογική δομή (ιεραρχία και ιδιότητες), μέχρι ακόμα και το φυσικό σχεδιάγραμμα. Τα ενεργά (προγραμματικά) στοιχεία ενός συγκεκριμένου εγγράφου ή ενός είδους εγγράφου, όπως οι υπερσύνδεσμοι ή το προσχεδιασμένο άνοιγμα και κλείσιμο, εφαρμόζονται από τις συμπεριφορές, οι οποίες (όπως προαναφέρθηκε) είναι ένας κώδικας του προγράμματος με πλήρη πρόσβαση στο περιεχόμενο των εγγράφων. Οι ιδιαίτερες συμπεριφορές που εφαρμόζονται σε ένα έγγραφο ή ένα είδος εγγράφου απαριθμούνται στη βάση των προτύπων της γλώσσας XML.

Το υπόλοιπο τμήμα αυτής της μελέτης υποδεικνύει εκείνες τις αρχιτεκτονικές έννοιες που εξετάζουν τις συγκεκριμένες πτυχές του προβλήματος ψηφιακής συντήρησης των μορφότυπων εγγράφων.

Σημείωση : Το πολυχρηστικό πρότυπο γεννήθηκε ως ένα project/μελέτη-εργασία μιας διατριβής στο Πανεπιστήμιο του Μπερκλεϊ , και ο δημιουργός του έχει μετακινηθεί από τότε στο πανεπιστήμιο του Λίβερπουλ.

3.1 Προσαρμογείς Μέσων

Οι νέοι μορφότυποι εγγράφων υποστηρίζονται από τους προσαρμογείς μέσων, οι οποίοι είναι τμήματα του κώδικα που μεταφράζουν τους υπαρκτούς/σταθερούς μορφότυπους με τα δεδομένα των εγγράφων σε δομές δεδομένων τρέχοντα χρόνου, πρωτίστως δε το δέντρο εγγράφων. Οι τρέχοντες/τωρινοί εφαρμοσμένοι προσαρμογείς μέσων περιλαμβάνουν, μεταξύ άλλων, το PDF, HTML, σκαναρισμένα έγγραφα των δύο OCR (οπτική αναγνώριση χαρακτήρων) - προτύπων, των σελίδων Unix, του TeX DVI, του ASCII, και του Apple II AppleWorks- επεξεργασία κειμένου. Οι προσαρμογείς μέσων ενσωματώνουν την μορφότυπο-συγκεκριμένη γνώση που τους αναλύει και είναι υποχρεωμένα να εξαλείψουν οποιαδήποτε ανάγκη για περαιτέρω αναφορά στο συγκεκριμένο/σταθερό bitstream. Αυτό συνεπάγεται διόρθωση του μορφότυπου οπουδήποτε απαιτείται (π.χ εξαναγκάζει τη HTML να μεταβάλλεται/τροποποιείται σε ένα DTD) και να παρουσιάζει το υπόλοιπο του συστήματος με ομοιόμορφες μονάδες λέξεων, οι οποίες μπορεί να απαιτήσουν/χρειαστούν τον διαχωρισμό/τεμαχισμό των γραμμών στο ASCII ή τη συγκόλληση των τεμαχισμένων τμημάτων της λέξης σε PDF ή TeX DVI.

Το σύστημα του πυρήνα (το βασικό κεντρικό σύστημα) δεν «ενσωματώνει» κανέναν προσαρμογέα μέσων επίσημα, αν και μερικοί δημοφιλείς από αυτούς συμβαίνει να συλλέγονται με τη συνηθισμένη διανομή. Το σύστημα του πυρήνα συνδέει μόνο ένα τύπο «μίμησης» ή ένα επίθεμα ενός τύπου αρχείου σε ένα μέρος/κομμάτι του κώδικα. Το σύστημα παρέχει όλα τα σύγχρονα προσβάσιμα και ελεγχόμενα έγγραφα γενικότερα. Επειδή δεν υπάρχει καμία διάκριση μεταξύ των ξεπερασμένων μορφότυπων εγγράφων και εκείνων που είναι σε χρήση, οι ξεπερασμένοι μορφότυποι εγγράφων είναι τόσο ενεργητικοί/ζωτικοί όσο εκείνοι σε τωρινή χρήση.

Οι προσαρμογείς μέσων διαβάζουν άμεσα τους πρωταρχικούς συγκεκριμένους/σταθερούς μορφότυπους των δεδομένων των εγγράφων. Αυτό αποφεύγει ένα πρόβλημα μετατροπής στο οποίο τα σφάλματα/οι παγίδες του λογισμικού ή οι προσεγγίσεις σε ένα στάδιο υποβιβάζουν κατά συρροή την ποιότητα. Τα σφάλματα/οι παγίδες του λογισμικού, παρότι είναι πάντα ανεπιθύμητα, είναι πιο καλόβολα/περιεκτικά στους προσαρμογείς μέσων, επειδή μόλις καθοριστούν/επανορθωθούν, όλες οι επόμενες όψεις (δηλ. αυτό που

βλέπουμε) και άλλες χρήσεις διορθώνονται αυτόματα. Με τον ίδιο τρόπο, οι επιμέρους (μερικές-όχι πλήρης) εφαρμογές των μορφότυπων, όπως εκείνες που γίνονται προσεκτικά βήμα-βήμα με μεταστροφή στην ανακατασκευή τους, είναι επαυξανόμενα βελτιωμένες. Οποιαδήποτε πρόοδος μπορεί να διαδοθεί και να γίνει εκμεταλλεύσιμη αμέσως, χωρίς καθυστέρηση έως ότου επιτυγχάνεται η τελειότητα και οι όποιες βελτιώσεις μπορούν να διανεμηθούν καθώς επιτυγχάνονται.

Η ικανότητα να διαβαστούν οι πρωταρχικοί μορφότυποι εγγράφων διασυνδέεται με το σύστημα και επομένως είναι πάντα διαθέσιμη όταν απαιτείται . Με τη μετατροπή, ίσως οι συσκευές που χρησιμοποιήθηκαν, μετέτρεψαν όλα τα γνωστά έγγραφα σε ένα σωρό (σε μεγάλη ποσότητα), ίσως και από έναν τρίτο παράγοντα και τώρα ο χρήστης αντιμετωπίζει μια νέα περίπτωση στην οποία πρέπει να αναβιώσει αυτή τη φάση. Με τη συντήρηση του υλικού, το μουσειακό υλικό και το λογισμικό έχει μια γεωγραφική θέση και ακόμα κι αν είναι στο δίκτυο, μπορεί να μην είναι σε θέση να ανοίγει τα εύθραυστα, αναντικατάστατα εκθέματά του στην κάθε τυχαία πίεση που δέχεται από τα εκατομμύρια των ατόμων που σερφάρουν στον Παγκόσμιο Ιστό. Με ένα καθολικό σχήμα, το γεγονός ότι οι νέοι μορφότυποι μπορεί να είναι εύκολο να αναλυθούν δεν μας κάνει και κανένα ιδιαίτερο καλό εκτός και αν το έγγραφο έχει προεπεξεργαστεί.

Οι προσαρμογείς μέσων χρησιμεύουν ως οι λειτουργικοί ορισμοί των μορφότυπων εγγράφων. Όταν οι προσαρμογείς μέσων είναι το αποτέλεσμα της αντίστροφης εφαρμοσμένης μηχανικής, οι λειτουργικοί ορισμοί τους χρησιμεύουν επίσης ως οι αναμφισβήτητες/σπάνταρντ προδιαγραφές. Όταν οι προσαρμογείς μέσων βασίζονται σε διαφορετικές καθορισμένες προδιαγραφές, παραμένουν απαραίτητοι/σημαντικοί δεδομένου ότι δια φωτίζουν τις σκοτεινές γωνίες του πραγματικού κόσμου και τη (κατα)χρήση που υπάρχει έξω από το φως/πρίσμα των προδιαγραφών. Παραδείγματος χάριν, η HTML όπως συναντάται στον Ιστό δεν είναι σχεδόν ποτέ σωστή και απαιτεί συχνά κάποια ιδιαίτερη/σημαντική διόρθωση, όχι για (να ακολουθηθούν) οι προδιαγραφές της HTML σύμφωνα με τα δεδομένα του W3C's, αλλά για τον καθορισμό λειτουργίας που δίνεται από τον ίδιο Microsoft Internet Explorer. Στο πολυχρηστικό πρότυπο, αυτοί οι λειτουργικοί ορισμοί είναι μέρος ενός ζωντανού συστήματος, έτσι συντονίζονται πάντα και τηρούνται ενήμεροι.

Σε σύγκριση με τη μετατροπή, οι προσαρμογείς μέσω των μετακινούν το πρόβλημα της συντήρησης από μια σταθερή/συνεχόμενη εντριβή των δισεκατομμυρίων των εγγράφων προς τη διατήρηση ενός προσαρμογέα μέσω των ανά μορφότυπο. Η συντήρηση των μεμονωμένων εγγράφων μειώνεται με το να αντιγράφουν απλώς τα κομμάτια (τα bits , δηλαδή την ελάχιστη δυνατή πληροφορία τους).

3.2 Το Δέντρο Εγγράφων

Η αρχική δομή των δεδομένων είναι το δέντρο του εγγράφου, το οποίο αντιπροσωπεύει ολόκληρο το περιεχόμενο ενός εγγράφου (ως ενιαία κύλιση-του ποντικιού-, ή κατά μια σελίδα τη φορά). Η ιεραρχική δομή ενός εγγράφου αντικατοπτρίζεται άμεσα στην ιεραρχία των κομβικών συνδέσεων γονέα-παιδιού στο δέντρο, και όλες αυτές οι κομβικές συνδέσεις μπορούν να περιέχουν ιδιότητες. Για έγγραφα όπως τα SGML, XML και SVG, το δέντρο απεικονίζει άμεσα το αναλυτικό/διεξοδικό δέντρο του εγγράφου. Η HTML αντιπροσωπεύεται ομοίως, αλλά αφού πρώτα διορθωθεί σε DTD. Οι εσωτερικοί κόμβοι του δέντρου είναι δομικοί και τα φύλλα μένουν ικανοποιημένα (κείμενο, εικόνες bitmap). Όλοι οι κόμβοι έχουν από το σχεδιάγραμμα οριοθετημένα (τετράγωνα) σχήματα/κουτιά (συντεταγμένες και διαστάσεις), που εκεί βρίσκονται οι εσωτερικοί κόμβοι οι οποίοι περιέχουν το ενωτικό (τετράγωνο) σχήμα των παιδιών τους. Συνήθως η δομή και το σχεδιάγραμμα συμπίπτουν, αλλά μερικές φορές ένα ειδικό κλαδί στη ρίζα του δέντρου απαιτείται για να προσαρμόσει τις αποκλίσεις όπως είναι οι επιπλέον/επιπλέουσες εικόνες και οι πολλαπλές στήλες. Τα μεταδεδομένα είναι διαθέσιμα ως ιδιότητες στη ρίζα του δέντρου. Οι προσαρμογείς μέσω των μπορούν να εισαγάγουν νέους τύπους κόμβων όταν απαιτούνται, αντίθετα από το μοντέλο αντικειμένων των εγγράφων της HTML. Εντυπωσιακά, όλοι οι μορφότυποι εγγράφων που έχουμε δει ως τώρα ταιριάζουν άνετα σε ένα κοινό δέντρο εγγράφων, από το σταθερό-μορφότυπο του σκαναρισμένου εγγράφου και το PDF από τη μία πλευρά της ποικιλίας/γκάμας έως τις ευμετάβλητες σελίδες της HTML και του Unix στην άλλη πλευρά (της γκάμας).

Τα δέντρα εγγράφων των περίπλοκων μορφότυπων εγγράφων είναι «διακοσμημένα» με τις εκτάσεις. Οι υπερασύνδεσμοι και το στυλ/μορφοποίηση του φόντου/πηγής είναι και οι δύο τύποι έκτασης. Οι εκτάσεις παρέχουν τον έλεγχο φύλλο-προς-φύλλο (και μέσα στο φύλλο) της εμφάνισης : οικογένεια/είδος του φόντου, μέγεθος, ύψος, χρώμα πρώτου πλάνου και υποβάθρου, υπογράμμιση, πλάτος γραμμών και άλλα. Οι εκτάσεις ελέγχουν επίσης την αλληλεπίδραση, την αναφορά των κλειδιών/πλήκτρων καθώς τα πιέζουμε και τη δραστηριότητα του ποντικιού μέσα στην έκταση. Μια «χούφτα» των εκτάσεων επαναχρησιμοποιείται από τους προσαρμογείς μέσω των για πολλούς διαφορετικούς μορφότυπους εγγράφων.

Σημειώνουμε ότι το δέντρο εγγράφων δεν είναι ένα ενδιάμεσο σχήμα ή ένα καθολικό σχήμα. Αντίθετα από ένα ενδιάμεσο σχήμα/μορφότυπο, το δέντρο εγγράφων χρησιμοποιείται άμεσα για την εμφάνιση και τη συμπεριφορά εγγράφων, που συντηρούν την πλήρη έκφραση των εγγράφων. (Το δέντρο εγγράφων υιοθετεί έννοιες κοινές μεταξύ των μορφότυπων όπου είναι δυνατόν και μπορεί να χρησιμοποιηθεί για τη μετατροπή). Στις περιπτώσεις που οι καθολικοί μορφότυποι μπορούν να επεκταθούν ως ένα ενωτικό σχήμα των ενσωματωμένων μορφότυπων, το δέντρο εγγράφων προσαρμόζεται σε ένα μεμονωμένο μορφότυπο τη φορά, χωρίς επιπλέον λειτουργικό κόστος.

Υποστηρίζοντας καθολικά την συντήρηση, το δέντρο εγγράφου ανοίγει την πρόσβαση σε όλο το περιεχόμενο του εγγράφου. Η μετατροπή, που είναι ακόμα μακριά (λειτουργικά) από το να τα ανοίγει/μπορεί να τα εξετάσει όλα, τυπικά αποβάλλει χαρακτηριστικά τους ασυνήθιστους τύπους περιεχομένου . Η εξομοίωση κρύβει τα περιεχόμενα σε ένα αδιαπέραστο «κιβώτιο/κουτί». Στην πραγματικότητα, κατά έναν τρόπο το δέντρο εγγράφου είναι ανώτερο από τους αρχικούς/πραγματικούς συντάκτες και θεατές του λογισμικού για έναν μορφότυπο, επειδή εκείνο το λογισμικό πιθανώς δεν έδωσε την πρόσβαση σε άλλες εφαρμογές, τουλάχιστον όχι σε μια τέτοια ακραία/μεγάλη περιεκτικότητα του κειμένου, των εικόνων, της δομής, του ύφους, και του σχεδιαγράμματος.

Το δέντρο εγγράφου ενοποιεί την αντιπροσώπευση όλων των μορφότυπων εγγράφων και τους ανυψώνει σε ένα κοινό σύνολο/σχήμα σύγχρονων περιληπτικών/αφαιρούμενων εγγράφων. Παραδείγματος χάριν, το λειτουργικό σύστημα - και οι οριζόμενες/ συγκεκριμένες από την εφαρμογή κωδικοποιήσεις χαρακτήρων, συμπεριλαμβανομένων των διάφορων τρόπων που συναλλάσσονται με τα μεγάλα διεθνή σύνολα χαρακτήρων , όλα αυτά ομαλοποιούνται σε Unicode (ενιαίο κώδικα). Τα εργαλεία και οι υπηρεσίες στοχεύουν στις περιλήψεις/αφαιρέσεις και αυτόματα εργάζονται με όλους τους μορφότυπους. Μια εφαρμογή ανάλυσης εγγράφου έχει πρόσβαση στο περιεχόμενο, το ύφος, και το σχεδιάγραμμα, αδιάφορα εάν η πηγή ήταν ανιχνευμένο/σκαναρισμένο έγγραφο ή TeX DVI. Αυτή η μηχανή αναζήτησης στο προηγούμενο της τομέα/τμήμα που υιοθέτησε το TeX DVI θα μπορούσε επίσης να συλλέξει και τους υπερσυνδέσμους σε DVI για να τις προσθέσει στον crawler της .

3.3 Συμπεριφορές

Τα ενεργά (προγραμματικά) στοιχεία ενός συγκεκριμένου εγγράφου ή ενός είδους εγγράφου, εφαρμόζονται από τις συμπεριφορές. Μια επέκταση του ανωτέρου(προαναφερθέντος) δέντρου εγγράφου είναι ένα παράδειγμα ενός τύπου συμπεριφοράς, ένας προσαρμογέας μέσω είναι μια συμπεριφορά και προσχεδιασμένο άνοιγμα καθώς και η κατάρρευση(κλείσιμο) εφαρμόζεται με έναν συνδυασμό συμπεριφορών.

Οι συμπεριφορές είναι ένας αυθαίρετος κώδικας του προγράμματος με πλήρη πρόσβαση στα περιεχόμενα των εγγράφων, το δίκτυο, και το δίσκο (υπό τον όρο του περιορισμού ασφάλειας, αλλά κανένα αρχιτεκτονικό περιορισμό). Οι συμπεριφορές μπορούν να είναι αυθαίρετα μεγάλες. Οι συμπεριφορές μπορούν αυθαίρετα να επεξεργάζονται το δέντρο εγγράφου. Ο μοναδικός περιορισμός σε μια συμπεριφορά είναι ότι εμμένει σε μια ορισμένη διεπαφή για την επικοινωνία με το σύστημα και κάποιες άλλες συμπεριφορές.

Για τη συντήρηση, οι συμπεριφορές αγκαλιάζουν πλήρως τις ενεργές και ιδιοσυγκρασιακές πτυχές ενός μορφότυπου εγγράφου. Δεν υπάρχει κανένας περιορισμός, για παράδειγμα, ποιο JavaScript μπορεί να έχει πρόσβαση σε ένα εγγενώς περιορισμένο επίπεδο διαμόρφωσης (του μορφότυπου). Παραδείγματος χάριν, το PDF καθορίζει ένα σύνολο τύπων σχολιασμού/υποσημειώσεων, όπως το μελάνι και το αποτύπωμα που κανένας από τους άλλους εφαρμοσμένους μορφότυπους εγγράφων δεν κάνει και με ένα σύνολο ιδιοτήτων που δεν μοιάζουν με οποιοδήποτε άλλο μορφότυπο εγγράφου. Κάθε τύπος σχολιασμών/υποσημειώσεων (μερικοί αλλά όχι όλοι όσοι εφαρμόζονται προς το παρόν) γίνεται ένας τύπος συμπεριφοράς και κάθε περίπτωση σχολιασμού/σημείωσης (γίνεται) μια περίπτωση συμπεριφοράς. Εάν ένα λογοτεχνικό υπερκείμενο χρειάζεται έναν νέο τύπο υπερσυνδέσμων με ειδικά χαρακτηριστικά γνωρίσματα, θα μπορούσε να τον εισαγάγει ως συμπεριφορά.

3.4 Οι Βάσεις / Τα Κέντρα (τα σημεία εστίασης)

Οι ιδιαίτερες συμπεριφορές που ισχύουν/εφαρμόζονται για ένα έγγραφο ή ένα είδος εγγράφου παρατίθενται/απαριθμούνται στις βάσεις του XML-μορφότυπου. Οι βάσεις χρησιμοποιούν τις ιδιότητες/γνωρίσματα της XML για να προσαρμόσουν τις γενικές συμπεριφορές (να περνούν ένα URL σε έναν υπερσύνδεσμο, παραδείγματος χάριν) κατά τον ίδιο τρόπο που οι γλώσσες προγραμματισμού υιοθετούν τις παραμέτρους για τις λειτουργίες. Οι βάσεις χρησιμοποιούν την ιεραρχία της XML για να εσωκλειστούν/τοποθετηθούν τα πιο σύνθετα δεδομένα που σχετίζονται με μια συμπεριφορά παραδείγματος χάριν, όταν ο χρήστης που συντάσσει ένα τύπο-σημείωσης σε υποσημείωση, η συμπεριφορά της σημείωσης σώζεται κάτω από τον πρώτο επίπεδο, και το περιεχόμενο της σημείωσης, δηλαδή το φόντο και τα χρώματά της, ακόμη και επιπλέον υποσημειώσεις σε εκείνη την υποσημείωση, εσωκλείονται/τοποθετούνται ιεραρχικά μέσα στη σημείωση.

Μερικές συμπεριφορές ισχύουν μόνο για ένα έγγραφο (π.χ. οι υποσημειώσεις), μερικές στο είδος (π.χ. ο περιγραφικός έλεγχος των χειρόγραφων σελίδων) και άλλες σε πολλαπλούς μορφότυπους (π.χ. διάφορα μενού – pop up – που μπορούν να στείλουν τη λέξη που επιλεχθεί με το δρομέα σε μια υπηρεσία καθορισμού της).

Οι συμπεριφορές που αναπτύχθηκαν για ένα μορφότυπο (ή - δεν αναπτύχθηκαν καθόλου- για κανέναν ιδιαίτερο μορφότυπο) μπορούν να συσχετιστούν με άλλους μορφότυπους μέσω των βάσεων και φέρνουν έτσι τις νέες ιδέες στους παλαιούς μορφότυπους. Αυτό δεν απαιτείται αυστηρά από τη συντήρηση (η πλήρης έκφραση του μορφότυπου είναι επαρκής/ικανοποιητική) αλλά είναι απαραίτητο για να φέρει τους παλαιότερους μορφότυπους έξω από το γκέτο τους, στην ισότητα με νεότερους, και τελικά, μάλλον κάποια μέρα οι σημερινοί μορφότυποι θα θεωρηθούν αρχαίοι επίσης. Οι χρήστες θέλουν να κάνουν υποσημειώσεις σε όλα τα έγγραφά τους, άσχετα εάν είναι PDF που έχουν τύπους σημειώσεων, ή ASCII ή WordStar (ή HTML!) οι όποιοι δεν έχουν, και οι βάσεις σχετίζουν/συνδυάζουν τη λειτουργία εκτός ζώνης και γι'αυτό το λόγο είναι ελεύθερες περιορισμών της εκφραστικότητας σε αυτούς τους μορφότυπους.

4 Πρακτικότητες

4.1 Συνειδητοποίηση της Πραγματικότητας

Εάν το πολυχρηστικό πρότυπο εγγράφων καθορίζει μια αρχιτεκτονική που ταιριάζει καλά στις ανάγκες της ψηφιακής συντήρησης εγγράφων, είναι αμέσως προφανές ότι θα απαιτήσει μια μεγάλης κλίμακας προσπάθεια εφαρμογής ώστε να αγκαλιάσει πλήρως τις εκατοντάδες ή τις χιλιάδες των μορφότυπων αρχείων. Ο μεγάλος αριθμός τους και το γεγονός ότι γενικά είναι σημασιολογικά ασυμβίβαστοι μεταξύ τους εγγενώς αναγκάζουν τη μεμονωμένη προσοχή και έτσι απαιτούν ιδιαίτερη προσπάθεια. Αλλά η προσπάθεια εφαρμογής δεν είναι τίποτα παραπάνω από τον Παγκόσμιο Εικονικό Υπολογιστή ή το CAMiLEON.

Ευτυχώς, η εργασία αυτή είναι ιδιαίτερα παραλληλισμένη με τις ανεξάρτητες ομάδες που εφαρμόζουν τους προσαρμογείς μέσω για τους διαφορετικούς μορφότυπους. Η πολυχρηστική αρχιτεκτονική καθορίζει τα απαραίτητα τεχνικά σημεία του συντονισμού, αλλά απ'την άλλη δεν επιβάλλει κανένα γραφειοκρατικό γενικό κόστος, και οι μεμονωμένες ομάδες μπορούν να επιλέξουν τους μορφότυπους της τοπικής σπουδαιότητας ή ενδιαφέροντος. Δεδομένου ότι όλοι οι προσαρμογείς μέσω είναι μέρος της ίδιας αρχιτεκτονικής, τα κοινά συστατικά/στοιχεία μπορούν να διαμοιράζονται, όπως γίνεται προς το παρόν για τη μορφοποίηση παραγράφων των διαφόρων/πολλαπλών φόντων και για τους υπερσυνδέσμους .

Ο στόχος μας είναι αρκετά πιο εύκολος από το συγκεντρωτικό σύνολο όλου του αρχικού λογισμικού εγγράφων. Η συντήρηση δίνει έμφαση στην εμφάνιση και τη συμπεριφορά και όχι στην ιδιαίτερα μεγάλη επεξεργασία ενός στοιχείου ενός συστήματος. Η επινόηση του μορφότυπου εγγράφου απαιτεί ιδιαίτερη διανοητική προσπάθεια, την οποία διαβάζουμε μόνο από μια προδιαγραφή. Χρησιμοποιούμε τη σύγχρονη τεχνολογία και τα εργαλεία, ενώ μερικά αρχικά συστήματα γράφτηκαν σε συμβολική γλώσσα (Assembly) για να τρέχουν στα 48 Kilobytes της μνήμης.

4.2 Συντήρηση της συντήρησης

Οποιαδήποτε στρατηγική συντήρησης θα πάρει τη συντήρηση/μονιμότητα για να την προσαρμόσει στις μελλοντικές τεχνολογίες, και το σύστημά μας συντηρείται με τον ίδιο τρόπο όπως ο UVC του Lorie. Στον UVC, οι εφαρμογές στοχεύουν σε έναν απλό πυρήνα και το πρόβλημα της συντήρησης μειώνεται στο να μεταφερθεί στον πυρήνα. Το Πολυχρηστικό εφαρμόζεται εξ ολοκλήρου σε γλώσσα Java, και ο δικός μας Παγκόσμιος Υπολογιστής είναι η εικονική μηχανή της Java (JVM = Java Virtual Machine). Η JVM είναι άμεσα ανάλογη με τον UVC δεδομένου ότι και οι δύο είναι εικονικές μηχανές λίγο πολύ στο επίπεδο της Συμβολικής γλώσσας (Assembly). Η JVM είναι κάπως πιο περίπλοκη, αλλά η αρχική εκτίμηση/θεώρηση δεν είναι η απόλυτη απλότητα αλλά μάλλον ένας πλήρης, αυστηρός καθορισμός (στοιχεία για τη γλώσσα προγραμματισμού Java μπορούμε να πάρουμε αναλυτικότερα από την εξής πηγή : Lindholm.T. – Yellin.F : The Java Virtual Machine Specification, 2nd edition, Addison-Wesley Longman Publishing Co., Inc. 1999) συνδεδεμένος με ένα "λογικό" επίπεδο της δυνατότητας επίτευξης της εφαρμογής. Ίσως δεν θα ήταν πολιτικά ευνοϊκό για την εταιρεία IBM να επικεντρωθεί σε/επιδείξει μια τεχνολογία Sun ως βασικό λίθο της στρατηγικής της συντήρησής της, αλλά το γεγονός ότι οι πολυάριθμες επιχειρήσεις έχουν εφαρμόσει συμβατές JVMs αποδεικνύει τη βιωσιμότητά/πρακτικότητα του.

Θα ήταν παράλογο να χτιστεί ένα μεγάλο σύστημα στη Συμβολική γλώσσα (Assembly) των εικονικών μηχανών. Μετακινούμενοι από την ικανοποίηση της βασικής απαίτησης ή απαίτησης κλειδί της αυτοσυντήρησης, προς τις εκτιμήσεις/θεωρήσεις της τεχνολογίας λογισμικού για την οικοδόμηση ενός μεγάλου συστήματος, πρέπει να επιλέξουμε μια γλώσσα υψηλού επιπέδου. (Η χρήση της JVM δεν υπονοεί και τη χρήση της ίδιας της γλώσσας Java , αφού πολλές γλώσσες προγραμματισμού μπορούν να συναθροιστούν/είναι συμβατές με τη JVM, όπως πολλά προγράμματα μπορούν να συνάγονται στους διαφορετικούς μικροεπεξεργαστές. Οι διαφορετικές ομάδες θα μπορούσαν να επιλέξουν διαφορετικές γλώσσες και να συνεργαστούν αποτελεσματικά) .

5 Μέλλον

Η πολυχρηστική αρχιτεκτονική είναι κατάλληλη για τη συντήρηση, αλλά σχεδιάστηκε αρχικά για χρήση σε έναν φυλλομετρητή . Θα μπορούσε να ωφεληθεί από τις θεωρήσεις των εμπειρογνομόνων για τη συντήρηση, για να εξασφαλίσει ότι η γενική προσέγγιση αγκαλιάζει πλήρως όλες τις ουσιαστικές λεπτομέρειες προτού να προωθηθεί μια μεγάλης κλίμακας προσπάθεια εφαρμογής.

Μια σημαντική υποέρευνα/υποεργασία θα είναι η συλλογή των προδιαγραφών των μορφότυπων εγγράφων. Αυτοί είναι σημαντικοί για την εγγενή θέση τους ως οι σκόπιμοι ορισμοί, και εξετάζοντας πόσο χρονοβόρα είναι η αντίστροφα εφαρμοσμένη μηχανική, είναι σημαντικό για τους μηχανικούς λογισμικού να έχουν εύκολη πρόσβαση σε αυτές τις προδιαγραφές.

Ο Wheatly.P [στο βιβλίο του : Έρευνα και αξιολόγηση των πηγών πληροφόρησης για τους μορφότυπους αρχείων και την τεκμηρίωση του λογισμικού (2003)] έχει καταχωρήσει/καταλογογραφήσει πολυάριθμα βιβλία και ιστοσελίδες και στις έρευνες/εργασίες που έχει κάνει υποστηρίζει ότι μπόρεσε και συνέλλεξε πολλούς τύπους διάφορων μορφότυπων αρχείων. Οι συνοδευτικές υποέρευνες/υποεργασίες θα πρέπει να συλλέξουν με τη σειρά τους τις εφαρμογές, οι οποίες είναι οι λειτουργικοί ορισμοί των μορφότυπων, και δείγματα εγγράφων για τους υπεύθυνους για την ανάπτυξη.

Αναμφισβήτητα η παρούσα τεχνολογία θα πρέπει να γενικευτεί και να αναδιαμορφωθεί και ήδη μια περιοχή στην ιδιαίτερη περιοχή είναι εμφανής. Τα έγγραφα συναντιόνται συχνά στα «περιτυλίγματα» των διάφορων ειδών, μερικές φορές για τη συμπίεση (όπως τα αρχεία zip) και μερικές φορές στα εικονικά συστήματα αρχείων (όπως στη δομημένη αποθήκευση που χρησιμοποιείται από τις εφαρμογές γραφείου-office- της Microsoft). Ένα υπόστρωμα των κατατμητών των εγγράφων θα χρειαστεί να αναλύσει αυτές τις δομές ώστε να παρέχει την πρόσβαση μέσα στα έγγραφα.

Αυτή η μελέτη έχει επικεντρωθεί στα ψηφιακά έγγραφα, αλλά υπάρχουν πολλοί ακόμα άλλοι τύποι μέσων (μερικοί των οποίων ενσωματώνεται στα έγγραφα) που έχουν ανάγκη από συντήρηση, όπως τα επιστημονικά δεδομένα, ο ήχος, τα μουσικά αποτελέσματα, τα βίντεο, πολυμέσα (όπως η εφαρμογή flash της Macromedia) και προγράμματα επιλογών/μενού των DVD, είναι μερικά μόνο από τα αναφερόμενα. Είναι ασαφές εάν όλοι αυτοί οι τύποι μπορούν να

προσαρμοστούν στο πλαίσιο μιας κοινής αρχιτεκτονικής. Αλλά το πολυχρηστικό πρότυπο έχει καταδείξει ήδη τη δυνατότητα εφαρμογής του σε ποικίλα έγγραφα που περιέχουν κείμενο, εικόνες, διανυσματικές γραφικές παραστάσεις, και τον προγραμματικό χειρισμό ενός δέντρου εγγράφων. Ακόμη και αν περιορίζεται στην κατηγορία μερικών εκατοντάδων μορφότυπων, τα δισεκατομμύρια των περιπτώσεων εγγράφων διεκδικούν μια αξίωση για κάποια κοινωνική αξία.

6 Συμπέρασμα

Σε σύγκριση με τις υπάρχουσες προσεγγίσεις στην ψηφιακή συντήρηση εγγράφων, το πολυχρηστικό πρότυπο εγγράφου προσφέρει ένα βήμα προς τα εμπρός. Έναντι της μετατροπής, το αρχικό έγγραφο παραμένει τέλεια συντηρημένο. Έναντι της εξομοίωσης, το περιεχόμενο του εγγράφου είναι εύκολα διαθέσιμο. Έναντι στον UVM του Lorie, το πολυχρηστικό καθορίζει την αρχιτεκτονική υψηλού επιπέδου που είναι απαραίτητη για τους μηχανικούς λογισμικού, και η εφαρμογή του πολυχρηστικού προτύπου σε έναν αριθμό σύνθετων και ξεπερασμένων μορφότυπων εγγράφων αποδεικνύει τη δύναμη της αρχιτεκτονικής και την ασυμβίβαστη καταλληλότητα του για τη συντήρηση. Το πολυχρηστικό πρότυπο είναι ένα αποδεδειγμένο σχέδιο στο παρόν για το μέλλον της συντήρησης του παρελθόντος.

Ακολουθεί στην επόμενη σελίδα :

πίνακας 1.

Σύγκριση των επιλεγμένων συστημάτων που χρησιμοποιούνται για την ψηφιακή συντήρηση

	PDF	UVM	CAMiLEON	Multivalent
Καθορισμός	1990	2001	2001 (?)	1997 εφαρμόστηκε για τη συντήρηση το 2004
Επιδείξεις (τι εκφράζει , που έχει εφαρμογή)	Όλα όσα μπορούν να τυπωθούν	JPEG και JIF, bitmapped εικόνες (παραμένει PDF στηριζόμενο βασικά σε μια μετατροπή σε HTML	διάμετατροπή μεταξύ SVG, Draw, WMF διανυσματικών γραφικών	PDF , HTML , σκαναρισμένα έγγραφα , TeX DVI , UNIX σελίδες , Apple II Appleworks
Μέθοδος	Οδηγός εκτυπωτή συλλαμβάνει το εκτυπώσιμο ρεύμα εισροής(stream) ή εφαρμογή την παράγει άμεσα. Ο μορφότυπος υποστηρίζεται αιώνια.	διαβάζει τ' αρχικό bitstream με έναν διερμηνέα εγγράφου	διαβάζει τ' αρχικό bitstream με μια ενδιάμεση αντιπροσώπευση - μετατροπή σε άλλο μορφότυπο	διαβάζει τ' αρχικό bitstream και φτιάχνει δομές δεδομένων εκτελέσιμου χρόνου
Δυνάμεις - Αντοχές	συλλαμβάνει τις στατικές πτυχές όλων των μορφότυπων, είναι καλά δομημένο και σαφές	πιθανή έως πλήρης έκφραση της εμφάνισης και της συμπεριφοράς ενός εγγράφου	σε σύγκριση με άλλη μετατροπή, μόνο κατά ένα επίπεδο γίνεται υποβάθμιση της ποιότητας	πλήρως εκφράζει την εμφάνιση και τη συμπεριφοράς ενός εγγράφου
χρήση απ' άλλες εφαρμογές	Χρήση του Acrobat ή μιας βιβλιοθήκης τρίτου/άλλου	απροσδιόριστη	στιγμιαίες επιλογές της εφαρμογής για την παραγωγή του μορφότυπου	Ζωντανή χρόνο-εκτελέσιμη διασύνδεση (επίσης, ανοιχτό σε μετατροπές)
Αρχιτεκτονική Εγγράφου	Δίνει έμφαση στη γραφική εμφάνιση , η δομή είναι εκφραζόμενη αλλά όχι κοινή	απροσδιόριστη	ενδιάμεση αντιπροσώπευση (είτε ένα αδέξιο ενωτικό σχήμα όλων των μορφότυπων ή αφήνοντας έξω τους ιδιοσυγκρασιακούς)	πλήρως ανεπτυγμένη (προσαρμογείς μέσω με δέντρο εγγράφου, δομή –σχεδιάγραμμα –συμπεριφορές – προεκτάσεις – σφάλματα , σταθερά και ρευστά σχεδιαγράμματα)
(εφαρμοσμένη) μηχανική του Λογισμικού	ο μορφότυπος εγγράφου είναι καλά τεκμηριωμένος, Acrobat API, πολλές βιβλιοθήκες τρίτων	χαμηλού επιπέδου Συμβολική γλώσσα UVM (στη πράξη χρησιμοποιεί Java)	άγνωστη	καλά εξασκημένο σύστημα API , υψηλού επιπέδου γλώσσα (Java)
Συντήρηση	προωθεί τα οικονομικά (κέρδη) της Adobe	μεταφορά της UVM σε νέες μηχανές	αναπτύσσει νέους παραγόμενους μορφότυπους “δίνει αρχές μακροζωίας στο λογισμικό ”	μεταφορά της Java VM σε νέες μηχανές
μειονεκτήματα	όλα πρέπει να δείχνουν σαν PDF (σταθερό σχεδιάγραμμα, σελιδοποιημένα) χάνει την ιδιοσυγκρασία και τη συμπεριφορά	ανώριμη εφαρμογή: χωρίς αρχιτεκτονική εγγράφου, η UVM είναι πολύ χαμηλού επιπέδου για την ανάπτυξη	τα σημασιολογικά κενά μετατροπών μεταξύ των μορφότυπων τους υποβαθμίζουν ή χάνουν δεδομένα. Χάνεται η συμπεριφορά	ενδογενής διασύνδεση των εφαρμογών ή ανάπτυξη δικών του εφαρμογών (κανένας συμβιβασμός ως προς την ποιότητα του εγγράφου)

References

1. Adobe Systems, Inc. PDF as a Standard for Archiving, Adobe white paper.
<http://www.adobe.com/products/acrobat/pdfs/pdfarchiving.pdf>
2. Birrell, A. and McJones, P. Virtual Paper Web site, (1995–1997).
<http://www.research.compaq.com/SRC/virtualpaper/>
3. Brown, A. Preserving the Digital Heritage: Building a Digital Archive for UK Government Records, In Proceedings of Online Information. (2003)
<http://www.nationalarchives.gov.uk/preservation/digitalarchive/>
4. IBM. Digital Asset Preservation Tool Web site,
<http://www.alphaworks.ibm.com/tech/uvic?Open&ca=daw-flHts-120204>
5. International Standards Organization. ISO/CD 19005-1, Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF (PDF/A), November 31, (2003).
[http://www.aiim.org/documents/standards/ISO_19005-1_\(E\).doc](http://www.aiim.org/documents/standards/ISO_19005-1_(E).doc)
6. Janssen, W.C. and Popat, K. UpLib: a universal personal digital library system, In Proceedings of the ACM symposium on Document Engineering, (2003) 234–242
7. Lindholm, T. and Yellin, F. The Java Virtual Machine Specification, 2nd edition, Addison-Wesley Longman Publishing Co., Inc. (1999)
8. Lorie, R. Long term preservation of digital information, In Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (2001) 346–352
9. Marshall, C.C. and Golovchinsky, G. Saving Private Hypertext: Requirements and pragmatic dimensions for preservation. In Proceedings of ACM Hypertext (2004) 130–138 A No-Compromises Architecture for Digital Document Preservation 277
10. Mellor, P, Wheatley, P, Sergeant, D. "Migration on Request : A Practical Technique for Digital Preservation" ECDL (2002)
11. Meehan, J., Taft, E., Chernicoff, S., Rose, C., Karr, R. PDF Reference, fifth edition, (2004)
12. Messerschmitt, D.G. Opportunities for Research Libraries in the NSF Cyberinfrastructure Program, ARL Bimonthly Report 229 (2003).
<http://www.arl.org/newsltr/229/cyber.html>
13. Multivalent Web site. <http://multivalent.sourceforge.net>

14. The National Archives. PRONOM Web site.
<http://www.nationalarchives.gov.uk/pronom/>
15. Phelps, T.A. Multivalent Documents: Anytime, Anywhere, Any Type, Every Way User- Improvable Digital Documents and Systems, Ph.D. Dissertation, University of California, Berkeley (1998)
16. Phelps, T.A. and Wilensky, R. The Multivalent Browser: a platform for new ideas, In Proceedings of Document Engineering, (2001) 58–67
17. Rodig, P., Borghoff, U.M, Scheffczyk, J., and Schmitz, L. Preservation of digital publications: An OAIS extension and implementation, In Proceedings of the ACM Symposium on Document Engineering, (2003) 131–139.
18. San Diego Supercomputer Center. Persistent Archive Testbed (PAT).
<http://www.sdsc.edu/PAT/>
19. Wheatley, P. Survey and Assessment of Sources of Information on File Formats and Software Documentation (2003)
http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf
20. Wotzits Format? Web site. <http://www.wotsit.org/>