ΑΡΧΕΙΟΘΕΤΗΣΗ ΙΣΤΟΠΕΡΙΕΧΟΜΕΝΟΥ ΚΑΙ ΔΙΑΤΗΡΗΣΗ ΨΗΦΙΑΚΗΣ ΜΝΗΜΗΣ - Η ΕΜΠΕΙΡΙΑ ΤΟΥ ΟΠΑ

Μιχάλης Βαζιργιάννης
Καθηγητής Οικονομικού Πανεπιστημίου Αθηνών
και Έφορος Βιβλιοθήκης

Η παρούσα εισήγηση θα είναι περισσότερο προσανατολισμένη στην τεχνολογία.

Θα σας περιγράψω σήμερα μια καινοτομική προσπάθεια που γίνεται τα τελευταία λίγα χρόνια στη Βιβλιοθήκη μας, η οποία αφορά την αρχειοθέτηση των ιστοσελίδων του Πανεπιστημίου – διαθεσιμη στο http://archive.aueb.gr.

Απλώς πριν ξεκινήσω, ήθελα να τοποθετήσω την ομιλία μου μέσα σε ένα ευρύτερο πλαίσιο το οποίο έχει να κάνει με τα μεγάλης κλίμακας δεδομένα –τα λεγόμενα big data. Είναι φανερό –είναι πλέον κοινός στόχος– ότι η θάλασσα των δεδομένων μέσα στην οποία ζούμε, και τα οποία παράγονται με ιλιγγιώδεις ρυθμούς, περιέχει πολλές προκλήσεις και τεράστια προστιθέμενη αξία που μπορεί να εξαχθεί από αυτά τα δεδομένα.

Γι' αυτό έχει υπάρξει ενός είδους –ας πούμε– ομαδοποίηση τεχνικών από διαφορετικές επιστήμες, η οποία έχει μπει κάτω από την ομπρέλα που ονομάζεται data science, επιστήμη των δεδομένων, με την έννοια ότι, εάν μπορούμε να κάνουμε τη μεταφορά, την παρομοίωση, τα δεδομένα είναι ένας πετρελαιοφόρος ορίζοντας μέσα από τον οποίο μπορούμε να αντλήσουμε πολύτιμη πληροφορία. Και στο χώρο αυτό, επενδύονται τεράστια ποσά, και σε προϋπολογισμούς και σε εξοπλισμούς, και πολιτικές αποφάσεις λαμβάνονται πλέον με βάση τα δεδομένα. Και

βέβαια, και σε επίπεδο έρευνας και ανάπτυξης, υπάρχουν τεράστιοι πόροι οι οποίοι επενδύονται.

Τώρα να επικεντρωθούμε λίγο στο θέμα της ομιλίας το οποίο αφορά την αρχειοθέτηση ιστοσελίδων. Όπως ξέρουμε, ο παγκόσμιος ιστός είναι ήδη μια από τις κύριες πηγές πληροφόρησης, διάδοσης πληροφορίας, οικονομικών συναλλαγών, ανταλλαγής μηνυμάτων μέσα από τα κοινωνικά δίκτυα, κ.λπ. Το θέμα είναι ότι οι σελίδες, το περιεχόμενο αυτών των σελίδων αλλάζει πολύ συχνά και με μεγάλη ταχύτητα. Αυτό σημαίνει ότι το περιεχόμενο, όπως το βλέπουμε σήμερα σε μια ιστοσελίδα, αύριο-μεθαύριο μπορεί να είναι διαφορετικό και πιθανότατα δεν έχει αποθηκευθεί πουθενά. Και όπως είναι κοινός στόχος πλέον, η οικονομία, η κοινωνία, ο πολιτισμός εν μέρει, και σε αυξανόμενο μάλιστα ποσοστό, εξελίσσονται στον παγκόσμιο ιστό. Επομένως, το να φυλάσσουμε και να αρχειοθετούμε το αξιόλογο περιεχόμενο από τις ιστοσελίδες είναι μια υποχρέωση των κοινωνιών, των οργανωμένων κοινωνιών, και ήδη αυτό γίνεται σε αρκετές ανεπτυγμένες χώρες από τις Εθνικές Βιβλιοθήκες.

Επομένως, τα δεδομένα μπορεί να χαθούν εάν δεν τα αποθηκεύσουμε. Επίσης, υπάρχει και το θέμα ότι τα υπολογιστικά συστήματα, επίσης, μπορεί να έχουν βλάβες και να χαλάσουν, οπότε το περιεχόμενο το οποίο φιλοξενείται μπορεί να χαθεί. Επιπλέον, υπάρχει το θέμα των λεγόμενων επιθέσεων οι οποίες μπορούν να καταστρέψουν το περιεχόμενο ιστοσελίδων και θα πρέπει να αντιμετωπίσουμε και αυτό το φαινόμενο. Επιπροσθέτως, ένα άλλο θέμα, το οποίο βέβαια αφορά λίγο πιο πολύ το μέλλον αλλά έχει τη σημασία του, είναι αυτό που αναφέρουμε εδώ ότι, ιδίως, στις μικρές συσκευές, δεν κοιτάζουμε πλέον ιστοσελίδες αλλά κοιτάζουμε-δουλεύουμε με εφαρμογές, οι οποίες εφαρμογές δεν έχουν αυτή την υφή και την παρουσίαση πληροφορίας όπως οι ιστοσελίδες και έτσι δεν μπορούμε να πάρουμε την πληροφορία που θέλουμε να την αποθηκεύσουμε.

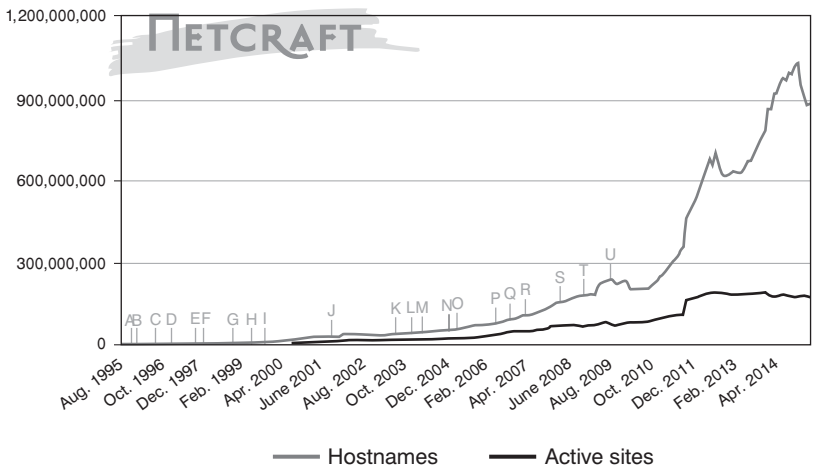
Επομένως, αναφύεται μια πρόκληση που έχει να κάνει με αυτή την τελευταία πρόταση εδώ ότι, στην ουσία, με την αρχειοθέτηση των ιστοσελίδων διασώζουμε την ιστορία της ανθρωπότητας, όπως αυτή εμφανίζεται στον παγκόσμιο ιστό.

Τώρα το Internet είναι ένα πάρα πολύ δυναμικό περιβάλλον. Έχω μερικούς αριθμούς οι οποίοι μας δείχνουν μέρος του δυναμισμού. Κάθε μήνα, οι ιστοσελίδες σε ποσοστό περίπου 8% αλλάζουν υπολογιστή φιλοξενίας. Δηλαδή, είναι σαν να λέμε ότι 8% των Βιβλιοθηκών αλλάζουν τοποθεσία. Πηγαίνουν σε άλλο κτήριο, θα λέγαμε, εάν μπορούμε να χρησιμοποιήσουμε αυτήν την αναλογία. Κάθε μέρα, ημερολογιακή, υπάρχουν περίπου 30.000 επιθέσεις σε ιστοσελίδες.

Επίσης, ένα περίπου 34% των βιομηχανιών οι οποίες διαθέτουν ιστοπεριεχόμενο δεν κάνουν ασφαλή αποθήκευση των αρχείων τους και συνεπώς κάποιο περιεχόμενο μπορεί να χαθεί. Επιπλέον, μόνο στις Ηνωμένες Πολιτείες, κάθε εβδομάδα, περίπου 140.000 σκληροί δίσκοι καταστρέφονται λόγω του ότι έχουν περιορισμενη διάρκεια ζωής. Συνεπώς, αναφύεται ένα πρόβλημα του πώς μπορούμε να έχουμε μια εγγύηση ότι το περιεχόμενό μας θα είναι ασφαλές.

Από την άλλη πλευρά, ο παγκόσμιος ιστός αυξάνεται με ιλιγγιώδεις ρυθμούς. Στο Διάγραμμα 1, βλέπουμε ότι, από το 1995 περίπου που ξεκίνησε η όλη ιστορία μέχρι τώρα, έχουμε μια σχεδόν επιθετική αύξηση τα τελευταία χρόνια των ονομάτων ιστοσελίδων. Ξέρετε, το όνομα μιας ιστοσελίδας είναι σημαντικό γιατί ουσιαστικά είναι σαν την ταμπέλα που βάζει κάποιος έξω από το κατάστημά του. Και βλέπετε ότι έχουμε εκατομμύρια σχεδόν ονόματα από τα οποία βέβαια λίγα, που αναπαρίστανται στο διάγραμμα με τη μαύρη γραμμή, είναι πλέον ενεργά. Αυτό σημαίνει ότι η διαφορά μεταξύ της γκρι και της μαύρης καμπύλης μάς δείχνει ιστοσελίδες οι οποίες ήταν ενεργές κάποια στιγμή, είχαν περιεχόμενο και αυτή τη στιγμή έχουν σταματήσει να ενημερώνονται, οπότε αποτελούν ένα αρχειακό υλικό. Μας ενδιαφέρει να τις αποθηκεύσουμε για να σώσουμε το περιεχόμενό τους;

ΔΙΑΓΡΑΜΜΑ 1
Evolution of the Web
Total number of websites (linear scale)

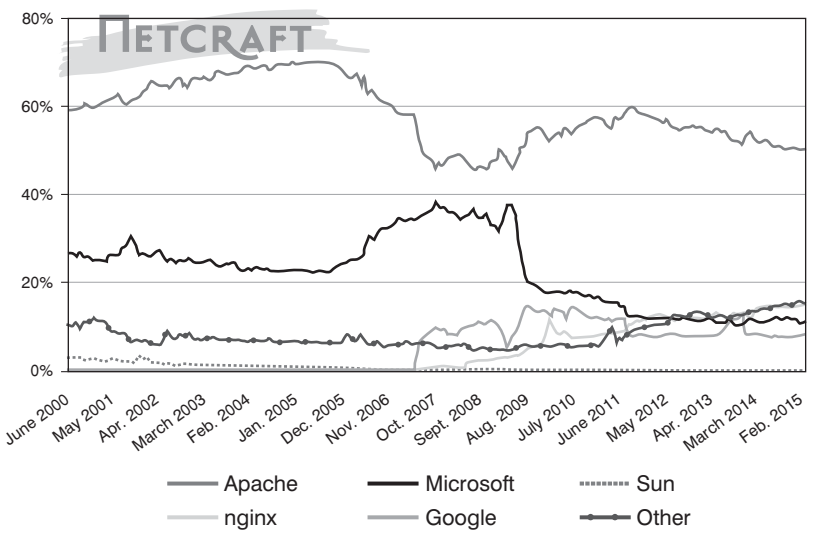


Πρέπει επίσης να εξετάσουμε πώς ένα σύστημα αρχειοθέτησης περιεχομένου που υπάρχει στον παγκόσμιο ιστό μπορεί να είναι χρήσιμο στην κοινωνία, στη βιομηχανία, στην οικονομία. Καταρχήν, θα μπορούσε κανείς να υποθέσει ότι ένας βιομηχανικός κλάδος μπορεί να βλέπει στις ιστοσελίδες των ανταγωνιστών και επειδή συνήθως οι ιστοσελίδες αντανακλούν τις εξελίξεις που γίνονται στις διάφορες εταιρείες και να μας δώσει μια καθαρή εικόνα των τάσεων που υπάρχουν στους διάφορους αυτούς τομείς. Επίσης, ιστορική ανασκόπηση της δομής και της αισθητικής των ιστοσελίδων μπορεί να μας δείξει το πώς άλλαξε ο παγκόσμιος ιστός σε επίπεδο σχεδίασης, σε επίπεδο διεπαφής με τον χρήστη και κ.λπ. Μπορεί να είναι χρήσιμο, επίσης, σε περιπτώσεις νομικών υποθέσεων όπου θέλουμε να θεμελιώσουμε ένα νομικό επιχείρημα, στηριγμένοι στο περιεχόμενο μιας ιστοσελίδας πριν από ένα, δύο, τρία, δέκα χρόνια, όπως για παράδειγμα στην περίπτωση παραβίασης πνευματικών δικαιωμάτων.

Το Διάγραμμα 2 μάς δείχνει την αναλογία των τεχνολογικών παικτών οι οποίοι αυτή τη στιγμή βρίσκονται στη βιομηχανία των servers, των εξυπηρετητών όπως λέμε, και του σχετικού λογισμικού. Από ό,τι βλέ-

πουμε, τη μεγάλη πλειοψηφία, το μέγιστο μερίδιο, το έχει η Apache που είναι ένας οργανισμός ανοικτού λογισμικού και το λογισμικό αυτό είναι αυτό το οποίο, κατά κόρον, χρησιμοποιείται για τη δημιουργία και φιλοξενία ιστοσελίδων σε διάφορους εξυπηρετητές. Τώρα, όπως ανέφερα και προηγουμένως, ένα από τα προβλήματα που μπορούν να υπάρξουν είναι, εάν δεν αποθηκεύουμε το περιεχόμενο των ιστοσελίδων ιστορικά, η πιθανή απώλεια του περιεχομένου, ενώ θα θέλαμε να το διατηρήσουμε το ιστορικό των ιστοσελίδων του και έτσι να προστατεύουμε και τη φήμη του. Είναι ένας είδος αρχείου των πάντων που γίνονται στον οργανισμό αυτόν.

ΔΙΑΓΡΑΜΜΑ 2
The Web market
Web server developers: Market share of active sites



Στην Ελλάδα, από όσο γνωρίζουμε, υπάρχει έλλειψη τέτοιων προσπαθειών και στη βάση αυτή, αυτών των κινήτρων, ξεκινήσαμε το 2010-2011, με βάση κάποια αρχική μας έμπνευση ερευνητική και στη συνέχεια με κάποια μικρή χρηματοδότηση από ένα έργο ψηφιακής σύγκλισης, την αρ-

χειοθέτηση των ιστοσελίδων του Οικονομικού Πανεπιστημίου, οι οποίες είναι πολλές σε αριθμό, όπως θα δούμε στη συνέχεια. Και βέβαια, δεν είναι μόνο το Οικονομικό Πανεπιστήμιο που το κάνει αυτό. Είναι μια διεθνής, όπως είπα, προσπάθεια –η κύρια προσπάθεια από την οποία εμπνέεται κανείς είναι το Internet Archive, που είναι μια διεθνής προσπάθεια στις Ηνωμένες Πολιτείες. Είναι μια μη κερδοσκοπική ψηφιακή βιβλιοθήκη, η οποία ιδρύθηκε το 1996.

Αυτή τη στιγμή, η συλλογή αυτή περιέχει ιστοπεριεχόμενο το οποίο ξεπερνάει τα 10 petabytes –μπορείτε να υπολογίσετε το μέγεθος της πληροφορίας– και αυξάνεται διαρκώς. Χρησιμοποιεί ένα λογισμικό που λέγεται Heritrix και θα δούμε αργότερα ποια είναι η σημασία του. Χρησιμοποιεί την τεχνολογία και το λογισμικό Petabox για την αποθήκευση αυτής της τεράστιας πληροφορίας και έχει, φιλοξενεί αυτή η πρωτοβουλία τέσσερα σημαντικά projects όπως το Wayback Engine, τις εικόνες του αρχείου της NASA, μια άλλη πρωτοβουλία που λέγεται Archive-it, και την Ανοικτή Βιβλιοθήκη.

Το Wayback Engine είναι μια τεχνολογία η οποία βοηθάει στην ανακατασκευή, στο να μας δείξει, δηλαδή, το πώς φαίνονταν παλαιότερα σελίδες που έχουμε αποθηκεύσει στο αρχείο μας. Δεν θα μπω σε τεχνικές λεπτομέρειες γιατί δεν αφορούν το ακροατήριο μας σήμερα, αλλά η κεντρική ιδέα είναι ότι είναι ένα λογισμικό ανοικτό το οποίο μας βοηθάει στην ανακατασκευή –αυτό που λέμε rendering– των ιστοσελίδων παλαιάς τεχνολογίας.

Το Archive-it είναι μια πλατφόρμα στην οποία μπορεί κάποιος οργανισμός να αποθηκεύει το ιστοπεριεχόμενό του και να το έχει διαθέσιμο στο μέλλον και μέχρι στιγμής, υπάρχουν 275 οργανισμοί οι οποίοι συμμετέχουν όπως οι βιβλιοθήκες των Ηνωμένων Πολιτειών, τα κρατικά τους αρχεία, και ούτω καθ' εξής. Είναι μια σημαντική προσπάθεια.

Επίσης, υπάρχει η προσπάθεια που λέγεται Open Library. Όλα αυτά κάτω από την ομπρέλα του Internet Archive, όπου στόχος αυτής της προσπάθειας είναι να υπάρχει μια ιστοσελίδα για κάθε βιβλίο το οποίο δημοσιεύτηκε κάποτε στην ιστορία. Και η πληροφορία αντλείται από τη Βιβλιοθήκη του Κογκρέσου, από το Amazon που είναι μια πολύ μεγάλη

πλατφόρμα η οποία ξεκίνησε πωλώντας βιβλία και φυσικά και από χρήστες οι οποίοι συνεισφέρουν βιβλία, τα οποία πιθανώς είναι παλαιότερα, ή και προφανώς είναι ανοιχτοί και σε άλλες συνεισφορές από βιβλιοθήκες εκτός του χώρου των Ηνωμένων Πολιτειών.

Και, τέλος, υπάρχει το International Internet Preservation Consortium, μία ομάδα οργανισμών οι οποίοι έχουν σαν στόχο τη διατήρηση του περιεχομένου του Internet και η οποία αριθμεί 48 μέλη από το 2014. Στόχος είναι η διατήρηση, η συλλογική διατήρηση, του ιστοπεριεχομένου.

Στο Χάρτη 1 βλέπουμε την κατανομή των προσπαθειών που γίνονται. Υπάρχει και στην Ελλάδα μια προσπάθεια δικιά μας, η οποία είναι καταχωρισμένη στην Wikipedia. Οι τεχνολογικές παράμετροι αυτού του εγχειρήματος είναι εντυπωσιακές για όσους και όσες από εμάς έχουν κατανόηση των τεχνικών παραμέτρων.

ΧΑΡΤΗΣ 1
International efforts
Countries that have made archiving efforts

Και ήθελα να δώσω πάρα πολύ γρήγορα και με πολύ γρήγορο ρυθμό την, σε γενικές γραμμές, αρχιτεκτονική του τρόπου λειτουργίας αυτών των

συστημάτων. Καταρχήν, πρέπει να αναζητήσουμε περιεχόμενο, δηλαδή να αποφασίσουμε ποιες σελίδες θα κατεβάσουμε και θα αρχειοθετήσουμε. Θα πρέπει να δημιουργήσουμε ένα ευρετήριο με αυτόματο τρόπο και να καθορίσουμε έναν τρόπο πρόσβασης στο αποθηκευμένο περιεχόμενο.

Το θέμα της συλλογής του περιεχομένου μπορεί να έχει διάφορες ενδιαφέρουσες παραμέτρους όπως, για παράδειγμα, το πώς θα διαλέξουμε ποιες ιστοσελίδες θα διατηρήσουμε, γιατί πρακτικά δεν μπορούμε να το κάνουμε για όλες. Επίσης, η συχνότητα με την οποία θα τις αποθηκεύουμε. Θα είναι κάθε εβδομάδα, κάθε μήνα, κάθε μέρα, κάθε χρόνο, το οποίο εξαρτάται από πολλές παραμέτρους. Έχουμε αυτή την παράμετρο που λέγεται politeness policy, δηλαδή, αν εμείς χτυπάμε τη σελίδα κάθε δευτερόλεπτο για να την αποθηκεύσουμε, δεν θα μπορεί να εξυπηρετήσει τις άλλες αιτήσεις, γι' αυτό θα πρέπει να είμαστε κατά κάποιο τρόπο ευγενικοί στο κάθε πότε και πόσο περιεχόμενο ζητάμε.

Και βέβαια, υπάρχει και η πολιτική του πώς κάνουμε παράλληλη αυτήν την τεχνική, δηλαδή πώς έχουμε πολλούς υπολογιστές να τρέχουν διαφορετικές τέτοιες εργασίες και να τελειώσουν πιο γρήγορα.

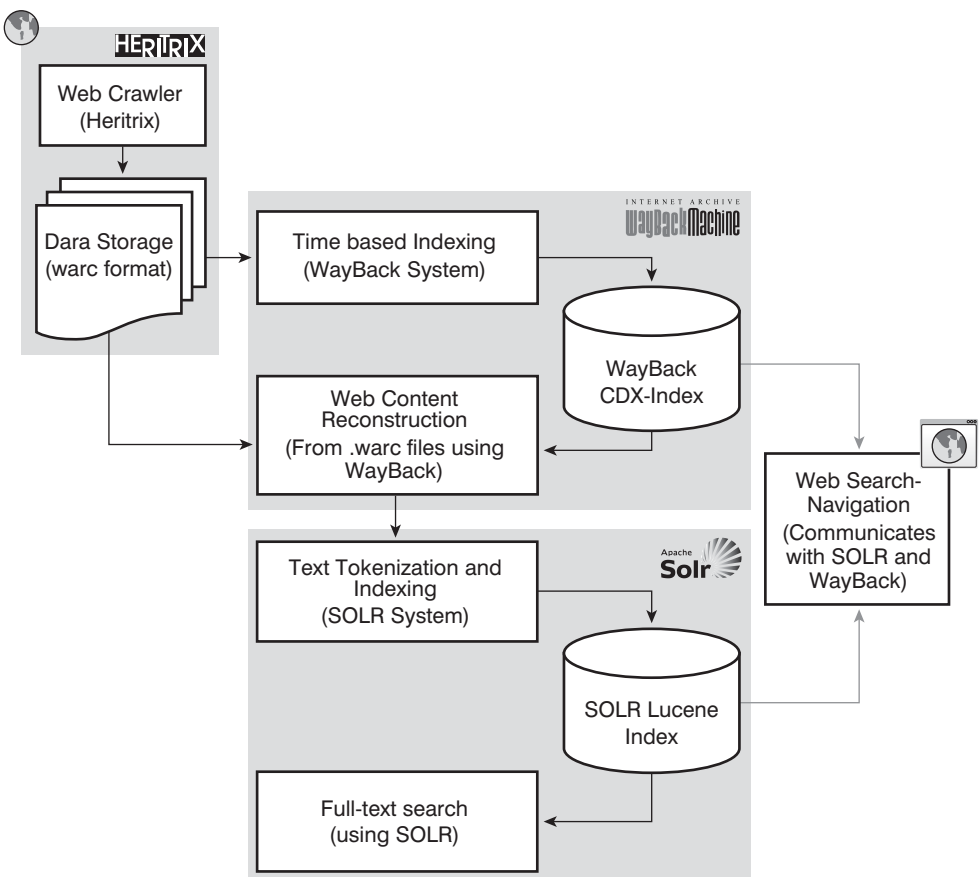
Οι στρατηγικές του πώς συλλεγουμε σελίδες είναι σε γενικές γραμμές οι εξής. Δηλαδή, μπορούμε είτε να πάμε εξαντλητικά ξεκινώντας από μια ιστοσελίδα και να συλλέξουμε όλα τα links στα οποία δείχνει αυτή η σελίδα και να συνεχίσουμε με αυτόν τον τρόπο που ονομάζεται αναζήτηση κατά πλάτος, είτε να πάμε σε αναζήτηση κατά βάθος, δηλαδή να ξεκινήσουμε από μια ιστοσελίδα, να μεταβούμε σε μία από αυτές τις οποίες δείχνει και μετά σε αυτήν να διαλέξουμε μία την οποία δείχνει και μετά με αυτήν να συνεχίσουμε την αναζήτηση κατά βάθος. Υπάρχουν πλεονεκτήματα και μειονεκτήματα.

Εδώ διακρίνεται κατά κάποιο τρόπο η τεχνολογική και η ερευνητική υποδομή που υπάρχει. Η μεγάλη επικεφαλίδα που μπορούμε να βάλουμε εδώ είναι επεξεργασία φυσικής γλώσσας και ανάκτηση πληροφοριών. Δεν θα σας απασχολήσω με περισσότερες τεχνικές.

Στο Διάγραμμα 3 φαίνεται η αρχιτεκτονική της δικιάς μας λύσης που υιοθετήσαμε στο Οικονομικό Πανεπιστήμιο, η οποία στηρίζεται σε ανοι-

κτό λογισμικό, εν πολλοίς παρόμοιο με τις αρχιτεκτονικές του Internet Archive. Το Heritrix είναι για τη συλλογή των δεδομένων. Το Wayback Engine είναι για την ανακατασκευή και να το ξαναδείξουμε στον χρήστη. Το Solr είναι η μηχανή με την οποία δημιουργούμε το ευρετήριο πάνω στο οποίο ψάχνουμε μετά, γιατί ο στόχος είναι να μπορούμε να ψάξουμε γρήγορα και αποδοτικά μέσα στην πληροφορία που συλλέγουμε. Και όπως είπα, είναι όλα ανοικτό λογισμικό. Και εδώ φαίνεται η λειτουργική αρχιτεκτονική.

ΔΙΑΓΡΑΜΜΑ 3
The AUEB web archiving case



Το Archive του Οικονομικού Πανεπιστημίου έχει πληροφορία από το 2010 και εντεύθεν κάθε μήνα και μπορούμε να ρωτήσουμε με διάφορους τρόπους με το url της ιστοσελίδας ή με λέξεις-κλειδιά οι οποίες αφορούν το περιεχόμενο των ιστοσελίδων και επίσης μπορούμε να θέσουμε και το χρονικό εύρος που μας ενδιαφέρει.

Τώρα, μερικά στατιστικά για το Archive του Οικονομικού Πανεπιστημίου (http://archive.aueb.gr) είναι ότι κάθε μήνα που κάνουμε τη συλλογή των δεδομένων, επισκεπτόμαστε όλες σχεδόν τις ιστοσελίδες του Οικονομικού Πανεπιστημίου, οι οποίες αθροίζουν σε περίπου 500.000 μοναδικές διευθύνσεις (URIs). Και έχουμε περίπου 500 διαφορετικούς εξυπηρετητές οι οποίοι έχουν πληροφορία για το διαδίκτυο κάτω από την ομπρέλα του Οικονομικού Πανεπιστημίου.

Υπήρχαν κάποια προβλήματα με υπολογιστές οι οποίοι μας στέλνανε ξανά στο ίδιο σημείο, όπως το forum των φοιτητών το οποίο το εξαιρέσαμε. Εν πάση περιπτώσει, τα δεδομένα τα οποία αυτή τη στιγμή διαθέτουμε είναι της τάξης των 27-32 gigabytes και είναι πληροφορία από το 2010 μέχρι σήμερα. Και, σε συμπιεσμένη μορφή, καταλαμβάνουν αυτό το χώρο. Δεν είναι τεράστιος, αλλά είναι σημαντικός. Μπορείτε να επισκεφθείτε και να κάνετε ερωτήσεις και να καταλάβετε πώς λειτουργεί αυτή η προσπάθεια στη διεύθυνση archive.aueb.gr. Σύντομα, θα έχουμε και ένα καινούργιο και καλύτερο interface.

Όπως έλεγα, ξεκινήσαμε από το 2010, αλλά νομίζουμε ότι αυτή η προσπάθεια είναι πιλότος για μια ίσως μεγαλύτερη και ευρύτερη προσπάθεια που χρειάζεται να γίνει στον ελληνικό χώρο. Κατά την άποψή μας, και ήδη είχαμε κάνει μια προσπάθεια το 2012, έχοντας αρχειοθετήσει μία και μοναδική φορά περίπου 1.400 διαφορετικές ιστοσελίδες του δημόσιου χώρου, περιλαμβάνοντας υπουργεία, δημαρχεία, πανεπιστήμια κ.λπ., αυτό έχει ιστορικό ενδιαφέρον, γιατί τότε που το κάναμε ήταν άλλη η διοικητική δομή της χώρας –που άλλαξε μετά τον Καλλικράτη– και έτσι έχουμε πληροφορίες για δήμους που δεν υπάρχουν πια, οπότε αναδεικνύεται μια αρχειακή ιστορική διάσταση.

Τώρα, επάνω σε αυτό το περιεχόμενο, εκτός από την αναζήτηση, μπορούμε να αναπτύξουμε και άλλες υπηρεσίες προστιθέμενης αξίας, όπως

για παράδειγμα να δημιουργήσουμε διάφορες ενδιαφέρουσες θεματικές συλλογές. Δηλαδή, να συλλέξουμε και να βάλουμε μαζί ιστοσελίδες οι οποίες είναι επάνω στην ίδια θεματική, να συστήσουμε ιστοσελίδες στους χρήστες μας που ψάχνουν οι οποίες είναι σχετικές με τις αναζητήσεις που κάνουν.

Υπάρχει και η ιδέα του λεγόμενου WikiMuseum, δηλαδή καθώς μιλάμε για αρχειοθέτηση κάποια από τα αντικείμενα που συλλέγουμε έχουν μια ευρύτερη αξία την οποία θέλουμε να αναδείξουμε σε ενός είδους online μουσείο, στο οποίο θα μπορούν να συμμετέχουν και οι χρήστες με τη δικιά τους συνεισφορά, αξιολόγηση, κ.λπ.

Τώρα υπάρχουν διάφορα θέματα όπως πόσο από το Internet, πόσο από τις ιστοσελίδες μπορούμε να καλύψουμε. Μπορούμε να κάνουμε το 100% κάθε χρονική στιγμή; Όχι προφανώς. Επομένως, θα πρέπει να πάρουμε κάποιες αποφάσεις όσον αφορά το τι καλύπτουμε και με τι συχνότητα.

Υπάρχει το θέμα του Deep Web, του βαθέος παγκόσμιου ιστού το οποίο έχει να κάνει με περιεχόμενο το οποίο βρίσκεται πίσω από φόρμες, πίσω από βάσεις δεδομένων που δεν είναι απλό περιεχόμενο, αλλά θα πρέπει να γίνουν ερωτήσεις για να το ανακτήσουμε. Και βέβαια, υπάρχουν και θέματα ιδιωτικότητας, copyrights, εάν μας επιτρέπει ο οργανισμός να κατεβάσουμε το περιεχόμενο και με τι άδειες κ.λπ.

Σε γενικές γραμμές, αυτή είναι η προσπάθεια που έχουμε κάνει.

Βιβλιογραφία

Pavalam, S., S. Raja, F. Akorli, and M. Jawahar. “A Survey of Web Crawler Algorithms”. IJCSI *International Journal of Computer Science Issues*, Vol. 8, Issue 6, No 1, November 2011.

Jaffe, Elliot and Kirkpatrick, Scott. “Architecture of the Internet Archive”. ACM. p. 11:1--11:10 2009.

Plachouras Vassilis, Chrysostomos Kapetis, and Michalis Vazirgiannis, “Archiving the Web sites of Athens University of Economics and Business”, in the 19th Greek Academic Library Conference (Nov. 2010).

Gomes, Daniel, João Miranda, and Miguel Costa. *A survey on web archiving initiatives.* Foundation for National Scientific Computing, 2011.

Udapure, Kale, Dharmik. “Study of Web Crawler and its Different Types”. *IOSR Journal of Computer Engineering.* Volume 16, Issue 1, Ver. VI (Feb. 2014).