Ανάπτυξη αποθετηρίων ανοικτής πρόσβασης: Ζητήματα και λύσεις

Νίκος Χούσος, Δέσποινα Χαρδούβελη, Παναγιώτης Σταθόπουλος, Κώστας Σταμάτης, Ηλίας Σταυράκης

{nhoussos; dxardo; pstath; stavrakis; kstamatis}@ekt.gr

Περίληψη

Η παρούσα εργασία αναφέρεται στην υλοποίηση δύο αποθετηρίων με αρκετά διαφορετικά χαρακτηριστικά του διεπιστημονικού ιδρυματικού αποθετηρίου του Εθνικού Ιδρύματος Ερευνών (Ήλιος) και ενός επικεντρωμένου στις ανθρωπιστικές επιστήμες και το ιστορικό / πολιτιστικό περιεχόμενο (Πανδέκτης). Δίνονται στοιχεία για το περιεχόμενο των αποθετηρίων και παρουσιάζονται επιλεκτικά διάφορες πτυχές της υλοποίησης της τεχνικής υποδομής των παραπάνω Αποθετηρίων όπως επεκτάσεις στο λογισμικό DSpace (διαχείριση καταλόγου καθιερωμένων ονομάτων, πρόσθετες λειτουργίες αναζήτησης, δυναμική αλλαγή γλώσσας, διασύνδεση αντικειμένων μέσω υπερσυνδέσμων κα.) Τέλος παρουσιάζονται τεχνικά θέματα που αφορούν στην ανάπτυξη και συντήρηση της απαραίτητης υποδομής των συστημάτων.

Λέξεις κλειδιά: αποθετήρια, ανοικτή πρόσβαση, περιεχόμενο, DSpace, έλεγχος καθιερωμένων ονομάτων, πολυγλωσσικότητα, αναζήτηση, lucene, virtualisation.

Εισαγωγη

Η ανάπτυξη αποθετηρίων ανοικτής πρόσβασης είναι ένα αντικείμενο στο οποίο παρατηρείται ιδιαίτερη αυξημένη δραστηριότητα από την ελληνική και διεθνή επιστημονική κοινότητα τα τελευταία χρόνια [1][2][3][4]. Το Εθνικό Κέντρο Τεκμηρίωσης συμμετέχοντας ενεργά σε αυτή τη διεθνή τάση προχώρησε στη δημιουργία ενός διεπιστημονικού ιδρυματικού αποθετηρίου που αφορά το σύνολο της ερευνητικής παραγωγής του Εθνικού Ιδρύματος Ερευνών («Ήλιος») καθώς κι ενός αποθετηρίου επικεντρωμένου στις ανθρωπιστικές επιστήμες και το ιστορικό / πολιτιστικό περιεχόμενο («Πανδέκτης»). Βασικός στόχος απετέλεσε η παρουσίαση και η ελεύθερη διάθεση μέσω διαδικτύου στο ευρύ κοινό, της ερευνητικής δραστηριότητας και του πνευματικού κεφαλαίου του Εθνικού Ιδρύματος Ερευνών αλλά και σημαντικών έργων της ελληνικής ιστορίας και της πολιτισμικής κληρονομιάς. Η εμπειρία και η τεχνογνωσία που αποκτήθηκε μέσω των συγκεκριμένων υλοποιήσεων αποθετηρίων στα θέματα ανοικτής πρόσβασης μπορεί να συμβάλει στη περαιτέρω εξέλιξη, διάδοση και υιοθέτηση υποδομών ανοικτής πρόσβασης στην ελληνική επιστημονική κοινότητα.

Η οργάνωση και λειτουργία των αποθετηρίων αποτελεί ένα σύνθετο σύνολο δράσεων που περιλαμβάνει ποικιλία ενεργειών, από τη συγκέντρωση και ψηφιοποίηση του υλικού, την επιλογή τύπων και μορφών περιεχομένου, το χειρισμό νομικών θεμάτων και πνευματικών δικαιωμάτων, τον προσδιορισμό σχημάτων μεταδεδομένων και την επικοινωνία με τους ερευνητές μέχρι την ανάπτυξη των κατάλληλων υποδομών λογισμικού και συστημάτων [1][3]. Το παρόν άρθρο, στην περιορισμένη έκταση του οποίου είναι αδύνατον να καλυφθούν σφαιρικά οι προαναφερθείσες δράσεις, παρουσιάζονται επιλεκτικά κάποια από τα ζητήματα που

αντιμετωπίστηκαν με έμφαση σε τεχνικά θέματα που αφορούν τα πληροφοριακά συστήματα και εφαρμογές υποστήριξης της λειτουργίας των αποθετηρίων.

ΔΡΑΣΕΙΣ ΑΠΟΘΕΤΗΡΙΩΝ ΤΟΥ ΕΚΤ

Η παρούσα ενότητα επιχειρεί να παρουσιάσει συνοπτικά τις δράσεις του ΕΚΤ που σχετίζονται με την ανάπτυξη των αποθετηρίων Ήλιος και Πανδέκτης.

Ήλιος

Ο «Ήλιος» αποτελεί το ιδρυματικό αποθετήριο του Εθνικού Ιδρύματος Ερευνών (ΕΙΕ), όπου συγκεντρώνεται σε ψηφιακή μορφή η ερευνητική δραστηριότητα του Εθνικού Ιδρύματος Ερευνών (ΕΙΕ) & προσφέρονται υπηρεσίες υποδοχής, οργανωμένης διαχείρισης και διάχυσης της επιστημονικής εκροής του οργανισμού σε ψηφιακή μορφή, αλλά και συλλογής, πρόσβασης, αποθήκευσης του ψηφιακού υλικού. Στη βάση των έξι ερευνητικών ινστιτούτων του Εθνικού Ιδρύματος Ερευνών (Ινστιτούτο Βιολογικών Ερευνών και Βιοτεχνολογίας, Ινστιτούτο Θεωρητικής και Φυσικής Χημείας, Ινστιτούτο Οργανικής και Φαρμακευτικής Χημείας, Ινστιτούτο Ελληνικής και Ρωμαϊκής Αρχαιότητας, Ινστιτούτο Βυζαντινών Ερευνών, Ινστιτούτο Νεοελληνικών Ερευνών), του Εθνικού Κέντρου Τεκμηρίωσης και των ειδικών μορφωτικών εκδηλώσεων του ΕΙΕ, δημιουργήθηκε υποδομή αποθετηρίου, με σκοπό την εισαγωγή και διαδικτυακή διάθεση των επιστημονικών δημοσιευμάτων και ερευνητικών εργασιών που παράγονται από τους ερευνητές τους.

Αυτή τη στιγμή μέσω του αποθετηρίου διατίθενται ελεύθερα στο Διαδίκτυο (στη διεύθυνση http://helios-eie.ekt.gr) 3.820 εγγραφές δημοσιευμάτων, εκ των οποίων περισσότερες από 1000 διαθέτουν ελεύθερα το πλήρες κείμενο. Το περιεχόμενο του Ήλιου ταξινομείται σε συλλογές σύμφωνα με τη διοικητική οργάνωση του ΕΙΕ. Οι κυριότεροι τύποι τεκμηρίων που περιλαμβάνονται στον Ήλιο είναι οι εξής: άρθρα επιστημονικών περιοδικών, μονογραφίες, βιβλία, κεφάλαια βιβλίων, περιοδικά, άρθρα σε περιοδικό και ημερήσιο τύπο, τεκμήρια επιστημονικών συνεδρίων και ημερίδων (πρακτικά, άρθρα, περιλήψεις, posters, παρουσιάσες), κείμενα εργασίας, εκθέσεις, εκπαιδευτικό υλικό, εγχειρίδια, οδηγοί, ενημερωτικά δελτία, video, κ.α**.** Στο αποθετήριο «ΗΛΙΟΣ» γίνονται αποδεκτές οι περισσότερες μορφές υλικού οι οποίες μετατρέπονται σε κατάλληλο μορφότυπο και εν τέλει διατίθενται κυρίως σε μορφή pdf και ppt.

Κατά την υλοποίηση του αποθετηρίου, πριν ακόμα από τη δημόσια λειτουργία του, δόθηκε έμφαση στη μαζική εισαγωγή αναδρομικού υλικού που έχει παραχθεί από τα Ερευνητικά Ινστιτούτα του ΕΙΕ. Αυτό κρίθηκε απαραίτητο κυρίως για να καταστεί εφικτή η διάθεση στο διαδίκτυο όσο το δυνατόν μεγαλύτερου μέρους της πλούσιου επιστημονικού έργου του ΕΙΕ σε σύντομο χρονικό διάστημα, γεγονός που αναδεικνύει το έργο του ΕΙΕ αλλά και για να αποτελέσει πυρήνα προσέλκυσης για τη συνεχή ενημέρωση του αποθετηρίου με νέο υλικό.

Το πρώτο βήμα για την αναδρομική εισαγωγή υλικού ήταν η καταγραφή και ο εντοπισμός των δημοσιευμάτων είτε σε έντυπη είτε σε ηλεκτρονική μορφή. Αυτό έγινε κυρίως μέσω της προσέγγισης των ιδίων των ερευνητών και των σχετικών υπηρεσιών του ΕΙΕ, αλλά και μέσω ηλεκτρονικών υπηρεσιών εκδοτών, βάσεων δεδομένων, πηγών στο Internet και έντυπων εκδόσεων περιοδικών. Αποτέλεσε μια ιδιαίτερα χρονοβόρα και επίπονη διαδικασία η οποία απαιτούσε την εξασφάλιση της συνεργασίας των ερευνητών αλλά και τη συστηματική αναζήτηση, ανάκτηση και οργάνωση περιεχομένου σε ηλεκτρονικές πηγές και πληροφοριακά συστήματα.

Αξίζει να αναφερθεί η διαφοροποίηση ανάμεσα στα θετικά και στα ανθρωπιστικά ινστιτούτα του ΕΙΕ, αφού όσο αφορά στα πρώτα, μεγάλο ποσοστό των πλήρων κειμένων κατέστη εφικτό να εντοπιστεί ηλεκτρονικά ενώ αντίθετα όσον αφορά στα ανθρωπιστικά Ινστιτούτα το συντριπτικό

μέρος του υλικού συγκεντρώθηκε σε έντυπη μορφή κατευθείαν από τους ερευνητές ή τις συναφείς υπηρεσίες των ινστιτούτων.

Τα επόμενα στάδια επεξεργασίας του υλικού μετά τη συγκέντρωσή του περιλάμβαναν ψηφιοποίηση, έλεγχο, μετατροπή σε μορφές αποδεκτές από το αποθετήριο και αποθήκευση με οργανωμένο τρόπο. Η παραπάνω διαδικασία, η οποία θα συνεχιστεί στο μέλλον για αναδρομικό υλικό που δεν έχει ακόμα εισαχθεί στο αποθετήριο, ενέχει ιδιαίτερες δυσκολίες λόγω του μεγάλου όγκου του υλικού, των ιδιαιτεροτήτων του υλικού (π.χ. χάρτες, εικόνες, έντυπα μεγάλης παλαιότητας, κ.α.), των απαιτούμενων διαδικασιών προσαρμογής της εικόνας του ψηφιοποιημένου αντικειμένου, ελέγχου και αποθήκευσης. Το κάθε τεκμήριο ψηφιοποιείται ολόκληρο αλλά και κατά άρθρο ή κεφάλαιο που περιλαμβάνει.

Συνολικά συγκεντρώθηκαν 1500 τεκμήρια σε έντυπη μορφή, (συμπόσια, συνέδρια, περιοδικά, τετράδια εργασίας, βιβλία, μονογραφίες, αυτοτελή δημοσιεύματα, άρθρα περιοδικών και πρακτικών συνεδρίων) που περιλαμβάνουν περισσότερα από 1700 κεφάλαια ή άρθρα). Από αυτά ψηφιοποιήθηκαν 900 τεκμήρια ολόκληρα αλλά και κατά άρθρο ή κεφάλαιο που περιλαμβάνουν, με σύνολο 74.896 σελίδες.

Πανδέκτης

Ο “Πανδέκτης - Ψηφιακός Θησαυρός Πρωτογενών Τεκμηρίων Ελληνικής Ιστορίας και Πολιτισμού" περιλαμβάνει σημαντικές ψηφιακές συλλογές ελληνικής ιστορίας και πολιτισμού, που συνοδεύονται από επιστημονική τεκμηρίωση των ανθρωπιστικών ινστιτούτων του Εθνικού Ιδρύματος Ερευνών. Μέσω του συγκεκριμένου αποθετηρίου διατίθενται ελεύθερα στο Διαδίκτυο (στη διεύθυνση http://pandektis.ekt.gr) πάνω από 35.000 εγγραφές και 20.000 ψηφιακά τεκμήρια. Τα περιεχόμενα των συλλογών, τα οποία προέρχονται από καταγραφές δεδομένων σε διάφορες μορφές από ερευνητές του ΕΙΕ, έχουν δομηθεί σε σχήμα βασισμένο σε qualified Dublin Core και είναι διαθέσιμα για αναζήτηση σε ενιαίο περιβάλλον.

Αυτή τη στιγμή ο Πανδέκτης αποτελείται από τις εξής συλλογές: Νεοελληνική Εικονιστική Προσωπογραφία, Ταξιδιωτική Γραμματεία, 15ος-19ος αιώνας, Τεκμήρια Ελληνικής Χαρτογραφίας, Μετονομασίες των Οικισμών της Ελλάδας, Αρχείο Εραλδικών Μνημείων του Ελλαδικού Χώρου, Έλληνες Ζωγράφοι μετά την Άλωση, 1450-1830, Οι Λειτουργοί της Ανώτατης, Μέσης και Δημοτικής Εκπαίδευσης (19ος αι.), Ελληνικός Τύπος του Εξωτερικού, Βιομηχανικές και Βιοτεχνικές Επιχειρήσεις στο Αιγαίο, Μοναστηριακά αρχεία: Έγγραφα Αγίου Όρους και Πάτμου, Αρχαίες ελληνικές και λατινικές επιγραφές της Άνω Μακεδονίας, της Αιγαιακής Θράκης και της Αχαΐας.

ΕΠΙΣΚΟΠΗΣΗ ΥΛΟΠΟΙΗΣΗΣ ΤΕΧΝΙΚΗΣ ΥΠΟΔΟΜΗΣ ΑΠΟΘΕΤΗΡΙΩΝ

Στην τρέχουσα ενότητα παρέχεται μια επισκόπηση των κυριότερων επιμέρους δράσεων που αφορούν την ανάπτυξη της τεχνικής υποδομής πληροφοριακών συστημάτων που είναι απαραίτητη για τη λειτουργία ψηφιακών αποθετηρίων επιστημονικού περιεχομένου και συγκεκριμένα: (α) την επιλογή πλατφόρμας λογισμικού αποθετηρίων, (β) τις αναγκαίες επεκτάσεις στην εν λόγω πλατφόρμα και (γ) τη δημιουργία της κατάλληλης υποδομής συστημάτων για την υποστήριξη τόσο των διαδικασιών ανάπτυξης όσο και της διάθεσης των αποθετηρίων.

Επιλογή λογισμικού

Για την επιλογή της πλατφόρμας λογισμικού ακολουθήθηκε μια αναλυτική διαδικασία αξιολόγησης των διαθέσιμων επιλογών. Διαπιστώθηκε από μια πρώτη έρευνα της κατάστασης διεθνώς ότι υπάρχουν αρκετά αξιόπιστα συστήματα ΕΛ/ΛΑΚ (Ελεύθερο Λογισμικό / Λογισμικό

Ανοικτού Κώδικα), οπότε η προσοχή μας επικεντρώθηκε σε αυτά. Μετά από ποιοτική αξιολόγηση με κύριο κριτήριο τη διεθνή αποδοχή και ωριμότητα εντοπίστηκαν ως επικρατέστερα τρία συστήματα: EPrints, Fedora/Fez, DSpace, για τα οποία πραγματοποιήθηκαν δοκιμαστικές εγκαταστάσεις και αναλυτικές δοκιμές με πραγματικά δεδομένα. Τελικά επιλέχθηκε το DSpace που, παρά τους περιορισμούς στην τρέχουσα (1.5.2) έκδοσή του (περιορισμένη υποστήριξη για σχήματα μεταδεδομένων, όχι τόσο σύγχρονες τεχνολογίες υλοποίησης, προβλήματα στη διαχείριση και τον έλεγχο πρόσβασης χρηστών), παρουσιάζει μια καλή ισορροπία ανάμεσα στην αξιοπιστία και την απλότητα δημιουργίας νέων εγκαταστάσεων με την υποστήριξη προηγμένων λειτουργιών.

Επεκτάσεις στο λογισμικό DSpace

Στην πορεία της δημιουργίας των αποθετηρίων Ήλιος και Πανδέκτης, παρουσιάστηκε η ανάγκη για σημαντικές επεκτάσεις του DSpace που απαιτούσαν αλλαγές στον κώδικά του. Οι συγκεκριμένες επεκτάσεις ήταν πέραν των συνηθισμένων παραμετροποιήσεων που είναι αναμενόμενες σε μια εγκατάσταση αποθετηρίου και για τις οποίες παρέχονται μηχανισμοί από το ίδιο το σύστημα (π.χ. διαφοροποιήσεις στην εμφάνιση, προσθήκη πεδίων μεταδεδομένων, προσδιορισμός και ευρετηρίαση αναζητήσιμων πεδίων).

Συνοπτικά οι κυριότερες επεκτάσεις που πραγματοποιήθηκαν είναι οι εξής:

Σύστημα διαχείρισης καταλόγου καθιερωμένων ονομάτων συγγραφέων.

Στα αποθετήρια, όπως και γενικά σε συστήματα που περιέχουν βιβλιογραφικές εγγραφές, εμφανίζεται το πρόβλημα των πολλών διαφορετικών γραφών του ονόματος του ίδιου συγγραφέα. Για τη μερική επίλυση του ζητήματος αυτού, που είναι ιδιαίτερα πολύπλοκο και δεν έχει αντιμετωπιστεί επαρκώς και σε διεθνές επίπεδο [5], έχει εφαρμοστεί ημι-αυτόματη διαδικασία που αναπτύχθηκε από το ΕΚΤ και βασίζεται εν μέρει σε πρωτότυπο μηχανισμό αυτοματοποιημένης παραγωγής καταλόγου καθιερωμένων όρων (authority file) για τα ονόματα των συγγραφέων που περιέχονται στο αποθετήριο. Η λειτουργία αυτή επιτρέπει την ενιαία γραφή του ονόματος του κάθε συγγραφέα στο αποθετήριο, ώστε να μπορούν να παρέχονται αξιόπιστα στους χρήστες υπηρεσίες πλοήγησης και αναζήτησης με βάση το όνομα συγγραφέα.

Πρόσθετες λειτουργίες αναζήτησης.

Στην πρότυπη εφαρμογή του DSpace οι υπηρεσίες αναζήτησης που προσφέρονται έχουν κάποιες βασικές ελλείψεις, οι οποίες έχουν καλυφθεί στις υλοποιήσεις των αποθετηρίων Ήλιος και Πανδέκτης. Οι κύριες βελτιώσεις που έχουν αναπτυχθεί αφορούν τη δυνατότητα αναζήτησης με χρονικά διαστήματα (ετών) και την ανεξαρτησία της αναζήτησης από τόνους και διαλυτικά (κάτι που προϋποθέτει στοιχεία υποστήριξης πολυγλωσσικότητας).

Δυνατότητα δυναμικής αλλαγής γλώσσας από το χρήστη

Η λειτουργία αυτή επιτρέπει στο χρήστη να αλλάξει την γλώσσα της γραφικής διεπαφής του αποθετηρίου σε πραγματικό χρόνο, με διατήρησή του στη σελίδα που βρίσκεται κατά την αλλαγή.

Προηγμένες λειτουργίες παρουσίασης για εικόνες

Δυνατότητες όπως μεγέθυνση/σμίκρυνση, προβολή σε πλήρες μέγεθος, περιστροφή και άλλες κρίνονται απαραίτητες για την πλήρη αξιοποίηση ψηφιακών εικόνων από ερευνητές αλλά και απλούς ενδιαφερόμενους/επισκέπτες του αποθετηρίου. Λαμβάνοντας υπόψη τα

παραπάνω, αναπτύχθηκε για το αποθετήριο Πανδέκτης ειδικό λογισμικό παρουσίασης εικόνων σε γλώσσα javascript που παρέχει τις προαναφερθείσες δυνατότητες, χωρίς να εξαρτάται καθόλου από το DSpace, άλλωστε μπορεί να χρησιμοποιηθεί και σε οποιοδήποτε ιστότοπο (όχι μόνο σε αποθετήρια). Το εν λόγω λογισμικό βασίζεται εν μέρει σε διάφορες εφαρμογές ΕΛ/ΛΑΚ και διατίθεται και το ίδιο από το ΕΚΤ ελεύθερα συμπεριλαμβανομένου του πηγαίου κώδικά του.

Προσαρμογή της παρουσίασης βάσει του περιβάλλοντος (context)

Στον Ήλιο και τον Πανδέκτη έχει ενσωματωθεί με συστηματικό τρόπο η δυνατότητα διαφοροποίησης της παρουσίασης σε επίπεδο συλλογής, εγγραφής αλλά και ψηφιακού τεκμηρίου (π.χ. διαφορετική εμφάνιση ανά συλλογή, διαφορετική ανάλυση εικόνων ανάλογα με περιορισμούς πνευματικών δικαιωμάτων). Οι απαραίτητοι κανόνες για την προσαρμογή της εμφάνισης εισάγονται με πολύ εύκολο τρόπο σε κώδικα Java. Η επέκταση αυτή έχει εφαρμοστεί στο μηχανισμό δημιουργίας γραφικών διεπαφών JSP UI του DSpace αν και μπορεί να χρησιμοποιηθεί ανεξάρτητα από αυτό. Το πλεονέκτημά της είναι ότι προσφέρει σημαντική ευελιξία αν και είναι πολύ απλή στην εφαρμογή της, σε αντίθεση με πιο εξεζητημένες λύσεις, όπως το Manakin [6] που εισάγει ιδιαίτερη πολυπλοκότητα προσφέροντας εν πολλοίς τα ίδια οφέλη. Σημειώνεται πως η συγκεκριμένη λειτουργία αφορά αποκλειστικά θέματα εμφάνισης/παρουσίασης στο ίδιο αποθετήριο υλικού διάφορων μορφών και απαιτήσεων. Για τη συνολικότερη αντιμετώπιση της ποικιλομορφίας του περιεχομένου και στο επίπεδο της διαχείρισης μεταδεδομένων απαιτούνται μεγαλύτερου εύρους λύσεις όπως αυτή που έχει εφαρμοστεί στην Ψηφιακή Βιβλιοθήκη Πέργαμος σε περιβάλλον αποθετηρίου Fedora για την αναπαράσταση και τη διαδικτυακή διάθεση συλλογών με υψηλό βαθμό ετερογένειας όσον αφορά τόσο τα μεταδεδομένα όσο και το ψηφιακό υλικό [7][8].

Διασύνδεση συγκεκριμένων συγγενών αντικειμένων μέσω υπερ-συνδέσμων.

Η επέκταση αυτή, αφορά αντικείμενα του αποθετηρίου μεταξύ των οποίων υπάρχει κάποια σχέση και επιτρέπει τη αμφίδρομη μετακίνηση του χρήστη μεταξύ των αντικειμένων μέσω υπερσυνδέσμων (π.χ. σύνδεση μεταξύ εικονογραφήσεων και των περιεχόντων βιβλίων στη συλλογή Ταξιδιωτικής Γραμματείας του Πανδέκτη).

Εξελιγμένη λειτουργία αυτό-αρχειοθέτησης.

Για το αποθετήριο Ήλιος έχει αναπτυχθεί φόρμα αυτο-αρχειοθέτησης πέρα από αυτή που παρέχει το DSpace, με δυνατότητες auto-complete για τους συγγραφείς και τους τίτλους περιοδικών Συγγραφέας,) καθώς και αυτόματης εξαγωγής και συμπλήρωσης στοιχείων όπου αυτό είναι εφικτό (π.χ. τα πεδία URL περιοδικού, ISSN, εκδότης συνάγονται από τον τίτλο περιοδικού).

Εισαγωγή λειτουργιών προσανατολισμένης πλοήγησης.

Η προσανατολισμένη πλοήγηση (faceted browsing) που δίνει στον χρήστη τη δυνατότητα «φιλτραρίσματος» στα μεταδεδομένα του αποθετηρίου με πολλά συνδυασμένα κριτήρια και οπτικοποίησης των αποτελεσμάτων με διάφορες μεθόδους (π.χ. σε χάρτη, σε χρονοδιάγραμμα, κλπ.) Η λειτουργία αυτή έχει ενσωματωθεί στη συλλογή των Εραλδικών μνημείων του Πανδέκτη. Για την υλοποίησή της χρησιμοποιήθηκε το λογισμικό ανοικτού κώδικα Exhibit [9] σε συνδυασμό με την υπηρεσία χαρτών Google Maps και αναπτύχθηκαν

επεκτάσεις για σύνδεση με αποθετήρια συμβατά με το πρωτόκολλο OAI-PMH, καθιστώντας το σύστημα ανεξάρτητο από το λογισμικό DSpace.

Όπως φαίνεται στην παραπάνω συνοπτική παρουσίαση, κατά την ανάπτυξη των επεκτάσεων έχει ληφθεί υπόψη ως μια από τις κύριες απαιτήσεις η ανεξαρτησία από την υφιστάμενη πλατφόρμα λογισμικού αποθετηρίων (το DSpace στις συγκεκριμένες υλοποιήσεις). Αξίζει να σημειωθεί πως κάτι τέτοιο δεν έχει καταστεί πλήρως εφικτό σε κάποιες περιπτώσεις, λόγω της απουσίας μιας πρότυπης προγραμματιστικής διεπαφής για λειτουργίες διαχείρισης περιεχομένου των αποθετηρίων. Σχετικές διεθνείς προσπάθειες έχουν καταλήξει στον ορισμό της διεπαφής SWORD [10] που υποστηρίζεται από τις κύριες πλατφόρμες αποθετηρίων, αλλά προς το παρόν αφορά μόνο την εισαγωγή περιεχομένου και όχι την ενημέρωση ή διαγραφή του, στοιχεία που είναι αναγκαία για τη διαλειτουργικότητα συστημάτων για αυτό-αρχειοθέτηση και χρήση καταλόγων καθιερωμένων όρων.

Στην Ενότητα 4 θα παρουσιαστούν με περισσότερες λεπτομέρειες οι επεκτάσεις που αφορούν την διαχείριση καταλόγου καθιερωμένων ονομάτων και της πολυγλωσσικότητας στην αναζήτηση.

Υποδομή συστημάτων

Οι σύγχρονες ψηφιακές βιβλιοθήκες/αποθετήρια απαιτούν εξελιγμένες τόσο τεχνολογικά όσο και οργανωτικά υποδομές συστημάτων για την αδιάλειπτη και αξιόπιστη παροχή υπηρεσιών προς τους χρήστες, την υποστήριξη μακροχρόνιας διατήρησης στο επίπεδο των συστημάτων καθώς και την ευελιξία στις διαδικασίες ανάπτυξη και παραγωγικής διάθεσης των υπηρεσιών. Οι παραπάνω ανάγκες δεν ικανοποιούνται από το μοντέλο της απλής εγκατάστασης μια σειράς πακέτων λογισμικού σε μονολιθικά συστήματα εξυπηρετητών, ούτε εξαντλούνται στην φάση ανάπτυξης και παραμετροποίησης του λογισμικού, αλλά απαιτούν μια ολοκληρωμένη προσέγγιση στα συστήματα υποδομής η οποία να αξιοποιεί τις διαθέσιμες τεχνολογικές λύσεις και τις διεθνώς αναγνωρισμένες καλές πρακτικές.

Με βάση αυτό το πλαίσιο, οι κύριες απαιτήσεις υποδομής συστημάτων για την υποστήριξη αποθετηρίων είναι η ασφάλεια και ακεραιότητα των δεδομένων, η υψηλότερη δυνατή διαθεσιμότητα των υπηρεσιών, η ευελιξία για ικανοποίηση συνεχώς διαφοροποιούμεων αναγκών με αποδοτικό τρόπο (χωρίς υπερ-διαστασιολόγηση), η υποστήριξη πολλών παράλληλων δράσεων ανάπτυξης και ποιοτικού ελέγχου για ανάπτυξη σε σύντομο χρονικό διάστημα καθώς και η μείωση της κατανάλωσης ενέργειας (υπολογίζεται ότι πάνω από το 2% τουλάχιστον της συνολικής κατανάλωσης ενέργειας στις ΗΠΑ π.χ. προέρχεται από συστήματα και υποδομές υπολογιστών [14]).

Για τους παραπάνω λόγους επιλέχθηκε η υποδομή συστημάτων του Ήλιου να έχει τα παρακάτω κύρια τεχνολογικά χαρακτηριστικά:

- Χρήση της τεχνολογίας virtualization [13], η οποία δίνει την δυνατότητα για τον ορισμό σε έναν εξυπηρετητή, πολλαπλών εικονικών μηχανών. Με αυτό τον τρόπο μπορούν να οριστούν «δεξαμενές» (pools) πόρων συστημάτων και αυτές να διαμοιράζονται ευέλικτα και δυναμικά σύμφωνα με τις πραγματικές ανάγκες κάθε εξυπηρετητή. Η υιοθέτηση του virtualization είχε ως αποτέλεσμα την σημαντική μείωση των απαιτούμενων φυσικών εξυπηρετητών καθώς και την αύξηση της διαθεσιμότητας υπηρεσιών αφού δεν υπάρχει καμία εξάρτηση ανάμεσα στην υπηρεσία και σε συγκεκριμένους φυσικούς εξυπηρετητές.
- Η διάθεση των αποθετηρίων σύμφωνα με το μοντέλο three tier δηλ. εγκατάσταση και λειτουργία σε διαφορετικούς εξυπηρετητές του τμήματος του web server, του application server, (δηλ. του DSpace στην συγκεκριμένη περίπτωση) και του εξυπηρετητή της Β.Δ. (PostgreSQL).

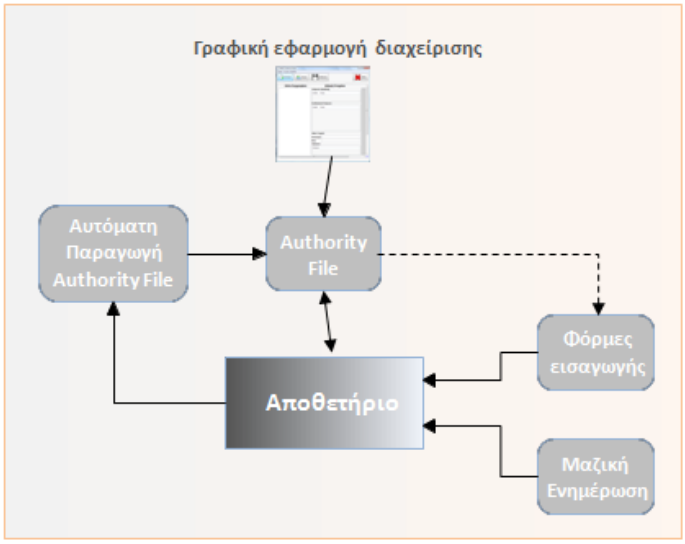
- Δημιουργία ξεχωριστών περιβαλλόντων εξυπηρετητών ανάπτυξης και δοκιμών από το περιβάλλον διάθεσης υπηρεσιών προς τους τελικούς χρήστες. Αυτός ο διαχωρισμός είναι στοιχειωδώς απαραίτητος όπως αναφέρεται και σε σειρά διεθνών προτύπων [11][12], αφού εξασφαλίζει την απομόνωση του περιβάλλοντος δοκιμών από αυτό το οποίο διατίθενται οι υπηρεσίες προς τους χρήστες. Στο χώρο των αποθετηρίων η πρακτική αυτή είναι επιτακτική, καθώς υπάρχει η ανάγκη για δημιουργία εγκαταστάσεων εκτός των παραγωγικών, για παράδειγμα, για έλεγχο λειτουργικότητας από τους ίδιους τους χρήστες και τους υπεύθυνους των συλλογών του αποθετηρίου ή/και για δοκιμαστική φόρτωση δεδομένων για ελέγχους από τους παραγωγούς περιεχομένου.
- Χρήση ΕΛ/ΛΑΚ, όπου είναι εφικτό, για λόγους ευελιξίας στη διαχείριση.
- Τη χρήση εργαλείων και συστημάτων παρακολούθησης επιδόσεων και υπηρεσιών καθώς και μηχανισμών έγκαιρης ειδοποίησης για προβλήματα που τυχόν ανακύπτουν σε συνδυασμό με μια εξειδικευμένη ολιγομελή ομάδα διαχειριστών συστημάτων.

Τα παραπάνω συνδυαζόμενα και με άλλες υποδομές όπως κεντρικό σύστημα λήψης αντιγράφων ασφαλείας, συστήματα ελέγχου φυσικής ασφάλειας, ταυτοποίηση διαχειριστών με ψηφιακά πιστοποιητικά, συνιστούν ένα κεντρικά διαχειριζόμενο περιβάλλον που παρέχει τα επιθυμητά χαρακτηριστικά για την λειτουργία των αποθετηρίων.

ΕΞΕΙΔΙΚΕΥΜΕΝΑ ΤΕΧΝΙΚΑ ΘΕΜΑΤΑ

Διαχείριση καταλόγου καθιερωμένων ονομάτων συγγραφέων

Η γενική διαδικασία για τον διαχείρισης καταλόγου καθιερωμένων ονομάτων συγγραφέων απεικονίζεται στην Εικόνα 1. Εξετάζοντας την περίπτωση ενός υπάρχοντος αποθετηρίου χωρίς οποιοδήποτε έλεγχο ονομάτων, το πρώτο βήμα είναι να παραχθεί αυτόματα ένα authority file σύμφωνο με το σχήμα MADS XML [16] μέσω κατάλληλου λογισμικού που έχει αναπτυχθεί από το ΕΚΤ και στηρίζεται σε μηχανισμό εντοπισμού ταυτότητας οντότητας μέσω του ονόματος (entity name disambiguation). Συγκεκριμένα, χρησιμοποιούνται τεχνικές χωρίς επίβλεψη (unsupervised) και ιδιαίτερα συσταδοποίηση (clustering). Το παραχθέν αρχείο είναι σύμφωνο με το πρότυπο MADS σε ό,τι αφορά τα καθιερωμένα ονόματα συγγραφέων και αποτελεί στη συνέχεια αντικείμενο επεξεργασίας από εξειδικευμένο προσωπικό, μέσω κατάλληλης εφαρμογής Java που έχει επίσης αναπτυχθεί εσωτερικά στο ΕΚΤ. Στην περίπτωσή μας, η χειρωνακτική επεξεργασία περιέλαβε και την προσθήκη επιπλέον πεδίων (π.χ., οργανισμός, θέση, στοιχεία επικοινωνίας) που συμπεριλαμβάνονται ως προαιρετικά στην προδιαγραφή MADS. Το προκύπτον διορθωμένο αρχείο χρησιμοποιείται για να ενημερώσει αυτόματα το αποθετήριο ‘Ηλιος σύμφωνα με την πολιτική μας που είναι να αντικαθίστανται τα εναλλακτικά ονόματα συντακτών με τα αντίστοιχα καθιερωμένα ονόματα. Για την ενημέρωση αυτή χρειάστηκε ειδική εφαρμογή που ενημερώνει τα περιεχόμενα του αποθετηρίου μέσω τεχνολογίας web services. Σημειώνεται πως καθιερωμένα ονόματα εφαρμόζονται μόνο για το ερευνητικό προσωπικό ΕΙΕ, για το οποίο είναι ρεαλιστικά εφικτός ο αξιόπιστος ποιοτικός έλεγχος των ονομάτων.



Εικόνα 1: Επισκόπηση διαχείρισης καταλόγου καθιερωμένων ονομάτων.

Η πρώτη εκτέλεση της προαναφερθείσας διαδικασίας οδήγησε στην δημιουργία του authority file μετά την αυτόματη διαδικασία παραγωγής και τον ανθρώπινο ποιοτικό έλεγχο. Στο τελευταίο αυτό στάδιο, μετρήθηκε ότι πραγματοποιήθηκαν από το προσωπικό του ΕΚΤ 40 διορθώσεις (μετακινήσεις ονομάτων από μια συστάδα σε άλλη), ενώ το αυτόματο πρόγραμμα εκτέλεσε 523 μετακινήσεις, κάτι που καταδεικνύει το κέρδος σε όγκο εργασίας από την εισαγωγή του εν λόγω προγράμματος. Πιο αναλυτικά, και με βάση τα αποτελέσματα της χειρωνακτικής παρέμβασης, η ακρίβεια της αντιστοίχησης ονομάτων σε συγγραφείς (precision – πόσα από τα εξευρεθέντα εναλλακτικά ονόματα ταξινομήθηκαν στο σωστό καθιερωμένο όνομα) έφτασε το 94.8%, ενώ η ανάκληση (recall – για πόσα από τα εναλλακτικά ονόματα που περιέχονταν στα δεδομένα εισόδου επιεύχθηκε αντιστοίχιση με καθιερωμένα ονόματα) ήταν 97.4%.

Όσον αφορά τις αναπροσαρμογές του περιεχομένου, κάποια πρέπει να εξετάσει τους ακόλουθους τρόπους ενημέρωσης ενός ιδρυματικού αποθετηρίου:

- Χειρωνακτικές αναπροσαρμογές από το προσωπικό βιβλιοθηκών και σε μερικές περιπτώσεις τους ίδους τους ερευνητές. Αυτός ο τύπος αναπροσαρμογής είναι ιδιαίτερα επιρρεπής σε τυπογραφικά λάθη και ασυμφωνίες, κάτι που αποφεύγεται σε κάποιο βαθμό με τις λειτουργίες auto-complete για ονόματα συγγραφέων που έχουμε εισάγει στη φόρμα αυτό-αρχειοθέτησης του ‘Ηλιου.
- Μαζικές αναπροσαρμογές, με εισαγωγή υλικού από διαθέσιμες πηγές όπως προϋπάρχουσες βάσεις δεδομένων, καταλόγους βιβλιοθηκών, ηλεκτρονικά περιοδικά ανοικτής πρόσβασης και άλλα συστήματα.

Ένα νέο authority file παράγεται αυτόματα μετά από κάθε μαζική αναπροσαρμογή batch και επίσης περιοδικά (π.χ., κάθε νύχτα ή κάθε εβδομάδα) για να ικανοποιήσει τις τροποποιήσεις που εκτελούνται μέσω της χειρωνακτικής εισαγωγής δεδομένων. Αυτό βελτιώνει την απόδοση και διευκολύνει τη συντήρηση των προηγουμένως εφαρμοσμένων χειρωνακτικών διορθώσεων. Η διαδικασία ενημέρωσης του αποθετηρίου εκτελείται πάντα μετά από την ολοκλήρωση, σε κάθε κύκλο αναπροσαρμογών, του χειρωνακτικού ποιοτικού ελέγχου.

Η προτεινόμενη προσέγγιση, σε σύγκριση με τον κατά κάποιο τρόπο σχετικό μηχανισμό που είναι σε διαδικασία ανάπτυξης για την επερχόμενη έκδοση 1.6 του DSpace [17], παρουσιάζει τις εξής διαφοροποιήσεις:

- Η προτεινόμενη από εμάς διαδικασία προβλέπει αμφίδρομη ανταλλαγή δεδομένων και λειτουργίες ενημέρωσης μεταξύ του συστήματος διαχείρισης του authority file και του DSpace, γεγονός που συντείνει στη βελτίωση της ποιότητας των δεδομένων και των δύο

συστημάτων (π.χ σε ακρίβεια για το αποθετήριο και κάλυψη για το authority file). Αντίθετα, στον υπό ανάπτυξη στοιχείο του DSpace υπάρχει ροή δεδομένων μόνο στην κατεύθυνση από το authority file προς το αποθετήριο.

- Το σύστημά μας καλύπτει και την πλήρη διαχείριση του authority file με βάση το πρότυπο MADS εκτός του αποθετηρίου με ειδική γραφική διεπαφή, κάτι που είναι εκτός πεδίου για το μηχανισμό του DSpace. Ο τελευταίος περιέχει και πολλές δυνατότητες διαμόρφωσης της πλατφόρμας του DSpace που δεν καλύπτονται στην υλοποίησή μας, αλλά δεν είναι και αναγκαία βάσει της διαδικασίας που προτείνουμε.

Επέκταση του Lucene για αναζήτηση στα ελληνικά

Η αναζήτηση στο DSpace υποστηρίζεται από το ΕΛ/ΛΑΚ λογισμικό Lucene, μια δημοφιλή μηχανή ευρετηρίασης και αναζήτησης υλοποιημένη σε γλώσσα Java. Κάποια παραδείγματα από το πλούσιο σύνολο λειτουργιών που παρέχονται από το Lucene είναι η σύνθετη αναζήτηση με boolean παραστάσεις, η ανεξαρτησία από πεζά-κεφαλαία και η «ασαφής» (fuzzy) αναζήτηση όπου επιστρέφονται αποτελέσματα που περιέχουν παρεμφερείς (όχι απολύτως ίδιες) φράσεις σε σχέση με το ερώτημα του χρήστη. Το πρόβλημα για τα ελληνικά αποθετήρια που βασίζονται στο DSpace είναι το γεγονός πως στην αναζήτηση δεν υπάρχει ανεξαρτησία από τον τονισμό λέξεων στην ελληνική γλώσσα. Συνεπώς αν ο χρήστης σε ένα ερώτημά του παραλείψει τους τόνους ή/και τα διαλυτικά (π.χ. αν δώσει τον όρο «βιολογια»), το σύστημα δεν θα του επιστρέψει κανένα αποτέλεσμα, αν ο συγκεκριμένος όρος περιέχεται σε εγγραφές του αποθετηρίου μόνο στην τονισμένη του μορφή.

Για την επίλυση του ζητήματος αυτού, χρειάστηκε η προσαρμογή του Lucene και ειδικότερα ενός στοιχείου του που ονομάζεται αναλυτής (analyzer). Κατά την ευρετηρίαση, ο αναλυτής χωρίζει το κείμενο σε αδιάσπαστα τμήματα/λέξεις (tokens), τα οποία αποτελούν τις καταχωρήσεις του ευρετηρίου. Η σημαντικότερη δουλειά του αναλυτή είναι η επεξεργασία των tokens που συνήθως συνίσταται στην στη μετατροπή όλων των γραμμάτων σε πεζά και σε αφαίρεση tokens που αντιστοιχούν σε συνήθεις λέξεις (stop words). Πιο εξελιγμένες εργασίες επεξεργασίας είναι η μετατροπή των τονισμένων χαρακτήρων σε άτονους και η αντικατάσταση των λέξεων με τις ρίζες τους (stemming). Κατά την αναζήτηση, ο αναλυτής εφαρμόζει τις ίδιες επεξεργασίες στο ερώτημα που θέτει ο χρήστης, ώστε τα tokens που προκύπτουν να είναι αντίστοιχα με αυτά που υπάρχουν στο ευρετήριο. Για παράδειγμα, η λέξη «Βιολογία» μετατρέπεται από το σύστημα σε «βιολογ» και με αυτή τη μορφή αποθηκεύεται στο ευρετήριο. Κάθε φορά που ο χρήστης περιλαμβάνει τον όρο αυτό (ή κάποιον με την ίδια ρίζα) στο ερώτημά του, ο αναλυτής τον μετατρέπει σε «βιολογ» ώστε να ταιριάζει με την καταχώρηση στο ευρετήριο. Συνεπώς ερωτήματα όπως «ΒΙΟΛΟΓΙΑ», «βιολογία», «Βιολογικός», «ΒΙΟΛΌΓΟΣ» επιστρέφουν τα ίδια αποτελέσματα.

Είναι προφανές ότι η επεξεργασία που πραγματοποιείται από τον αναλυτή είναι άμεσα συναρτώμενη με τη γλώσσα (π.χ. τα stop words, ο τονισμός, το stemming). Η προφανής προσέγγιση λοιπόν είναι η ύπαρξη ενός αναλυτή ανά γλώσσα. Στην τρέχουσα έκδοση του Lucene διατίθενται αναλυτές για πολλές γλώσσες, μεταξύ των οποίων και τα ελληνικά (η υλοποίηση του αντίστοιχου αναλυτή, που υποστηρίζει ανεξαρτησία από πεζά-κεφαλαία και από τονισμό έχει γίνει από τον Δρ Παναγιώτη Αστίθα, μηχανικό λογισμικού της ελληνικής εταιρίας EBS). Υπάρχει όμως ο περιορισμός ότι μόνο ένας αναλυτής μπορεί να είναι ενεργός στο σύστημα, άρα στην περίπτωση περιεχομένου σε πολλές γλώσσες οι παραπάνω λειτουργίες είναι διαθέσιμες για μία μόνο γλώσσα. Σημειώνεται ότι αυτό είναι πολύ σύνηθες σε αποθετήρια (για παράδειγμα στον Ήλιο υπάρχει περιεχόμενο σε τουλάχιστον τέσσερις διαφορετικές γλώσσες). Από την κοινότητα του DSpace έχει αναπτυχθεί ένας ξεχωριστός αναλυτής που πραγματοποιεί βασικές επεξεργασίες

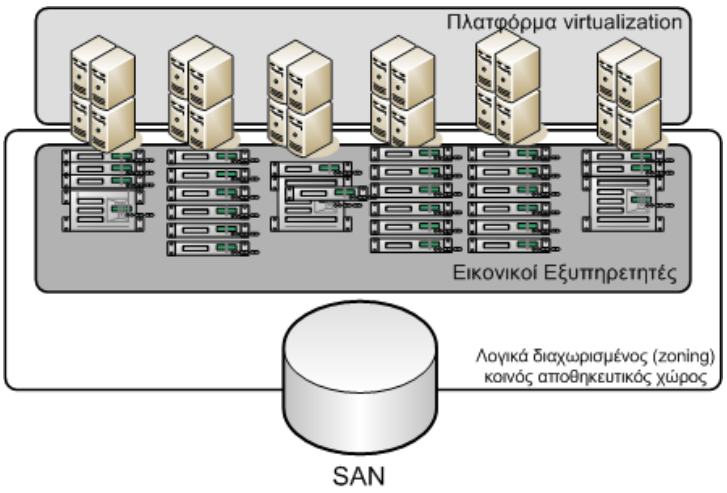
(π.χ. όσον αφορά θέματα του τονισμού) για διάφορες δυτικο-ευρωπαϊκές γλώσσες (αγγλικά, γαλλικά, γερμανικά) όχι όμως και για τα ελληνικά.

Ως λύση στα προαναφερθέντα προβλήματα, δημιουργήσαμε έναν παραμετρικό analyzer πολλαπλών γλωσσών με την χρήση των δυνατοτήτων που μας προσέφερε το Lucene. Ο analyzer αυτός μπορεί να δεχτεί ως διαμόρφωση, με την μορφή XML αρχείων, τα δεδομένα που χρειάζονται οι λειτουργίες που επιτελεί, δίνοντάς τους έτσι την δυνατότητα να επεξεργαστούν σωστά κείμενα από διαφορετικές γλώσσες. Στο αρχείο παραμετροποίησης του DSpace μπορεί εύκολα να ενεργοποιηθεί ο παραμετρικός αλγόριθμος ανάλυσης που περιγράψαμε αντί για τον αλγόριθμο ανάλυσης του DSpace (DSpace Analyzer). Στην συνέχεια μέσω αρχείων XML μπορεί κανείς να προσδιορίσει την σωστή κατάτμηση των κειμένων από τις γλώσσες της επιλογής του, π.χ. όσον αφορά stop words και αντικαταστάσεις γραμμάτων. Ο συγκεκριμένος αναλυτής, ο οποίος διαθέτει υποστήριξη για την ελληνική γλώσσα, έχει ενσωματωθεί στα αποθετήρια Ήλιος και Πανδέκτης, πρέπει όμως να τονιστεί ότι είναι ανεξάρτητος του DSpace, καθώς μπορεί να λειτουργήσει σε οποιοδήποτε σύστημα χρησιμοποιεί το Lucene ως μηχανή ευρετηρίασης / αναζήτησης.

Σημειώνεται επίσης ότι πιο απλοϊκές λύσεις στο προαναφερθέν πρόβλημα, όπως η επέκταση του ερωτήματος (query expansion) πάσχει από περιορισμούς στην κλιμάκωση, ειδικά όταν έχουμε ερωτήματα του τύπου ακριβούς φράσης (π.χ. όταν ένα ερώτημα με πολλές λέξεις περικλείεται από το χρήστη σε διπλά εισαγωγικά στο πλαίσιο κειμένου της μηχανής αναζήτησης).

Οργάνωση περιβάλλοντος ανάπτυξης και διάθεσης των αποθετηρίων

Η προσφορά των υπηρεσιών των αποθετηρίων γίνεται σε σημαντικό βαθμό μέσω της πλατφόρμας virtualization που έχει δημιουργήσει ο φορέας. Η πλατφόρμα απαρτίζεται από έξι εξυπηρετητές x86 με διπλούς τετραπύρηνους επεξεργαστές Intel Xeon και με συνολική διαθέσιμη μνήμη 176GB (τέσσερις με 32GB και δύο με 16GB). Οι εικονικές μηχανές εκτελούνται κατά βούληση στον εξυπηρετητή που έχει τους απαιτούμενους υπολογιστικούς πόρους. Η μετακίνηση των εικονικών μηχανών σε διαφορετικό φυσικό εξυπηρετητή γίνεται χωρίς καμία διακοπή της λειτουργίας τους μέσω της τεχνολογίας live migration. Για να είναι δυνατή η μετακίνηση αυτή, το σύνολο των αρχείων και οι ίδιες οι εικονικές μηχανές αποθηκεύονται σε κεντρικό σύστημα Storage Area Network. Στην πλατφόρμα εικονικών μηχανών έχουν αποδοθεί 8TB και ο κάθε φυσικός εξυπηρετητής συνδέεται με το σύστημα SAN με διπλές οπτικές οδεύσεις Fiber Channel 4Gbps και με δίσκους διαμορφωμένους σε RAID5. Παράλληλα συνδέονται με το δίκτυο με τέσσερις κάρτες δικτύου Gigabit Ethernet για την υποστήριξη του συνόλου των εκτελουμένων εικονικών μηχανών. Οι παραπάνω φυσικές συνδέσεις και εξαρτήματα χαρακτηρίζονται από τουλάχιστον N+1 υπερεπάρκεια. Η συγκεκριμένη υποδομή απεικονίζεται σχηματικά στην Εικόνα 2.



Εικόνα 2: Απεικόνιση της υποδομής συστημάτων βασισμένης σε τεχνολογίες virtualization και SAN.

Πάνω από αυτή την υποδομή δεν εκτελούνται μόνο οι υπηρεσίες των αποθετηρίων, αλλά και άλλες υποστηρικτικές υπηρεσίες που απαιτούνται για την λειτουργία και τον έλεγχο των αποθετηρίων, αλλά και επιπρόσθετες υπηρεσίες του φορέα. Πιο αναλυτικά εκτελούνται συνολικά 36 εικονικές μηχανές με λειτουργικό σύστημα Linux CentOS (έκδοση 5).

Το πλεονέκτημα είναι ότι οι παραπάνω υπηρεσίες μπορούν να μοιράζονται πόρους, αυξάνοντας ή μειώνοντας δεσμεύσεις πόρων ανά εικονικό εξυπηρετητή ανάλογα με τις ανάγκες. Σε περίπτωση που δεν είχε χρησιμοποιηθεί η τεχνολογία virtualization κάποιες υπηρεσίες θα ήταν υπερδιαστασιολογημένες ενώ κάποιες άλλες θα απαιτούσαν επιπλέον πόρους, με βάση τον δυναμικό και εξελισσόμενο χαρακτήρα τους. Υπογραμμίζεται ωστόσο ότι από την μέχρι στιγμής εμπειρία ο εξυπηρετητής που υποστηρίζει το DSpace παραμένει αυτός με τις μεγαλύτερες τεχνικές απαιτήσεις και συγκεκριμένα 8GB μνήμης και πολλά εκατοντάδες GB χώρου. Η πολιτική που ακολουθείται είναι της απόδοσης πόρων σύμφωνα με τις πραγματικές ανάγκες και όχι με βάση αρχική υπερδιαστασιολόγηση, μιας και αυτή η πολιτική γρήγορα θα εξαντλούσε τους πόρους με βάση υποθετικές ανάγκες χωρίς να υπάρχει η δυνατότητα για απόδοση τους σε επόμενο χρόνο και σε υπηρεσίες που πραγματικά τις απαιτούν.

Πέρα από την σημαντική αύξηση στην διαθεσιμότητα, μιας και σε περίπτωση αστοχίας ενός φυσικού εξυπηρετητή απλά οι εικονικές μηχανές επανεκινούνται μέσω του διαμοιραζόμενου αποθηκευτικού χώρου σε άλλο φυσικό εξυπηρετητή, επιτευχθηκε και σημαντική εξοικονόμηση ενέργειας, η οποία στην φάση της ανάπτυξης έχει υπολογιστεί ότι ήταν 45% σε σχέση με μια αντίστοιχη μη virtualized υποδομή [15].

Για την παρακολούθηση επιδόσεων και την διαχείριση του συστήματος χρησιμοποιούνται τα συστήματα ΕΛΛΑΚ Cacti και Nagios τα οποία παρακολουθούν δεκάδες συστήματα για εκατοντάδες παραμέτρους και υπηρεσίες εκτέλεσης.

Συμπερασματα – Μελλοντικες προοπτικες

Στην παρούσα εργασία παρουσιάστηκαν επιλεκτικά διάφορες πτυχές της υλοποίησης δύο αποθετηρίων με αρκετά διαφορετικά χαρακτηριστικά, ενός διεπιστημονικού ιδρυματικού που αφορά το σύνολο της ερευνητικής παραγωγής του Εθνικού Ιδρύματος Ερευνών (Ήλιος) και ένος επικεντρωμένου στις ανθρωπιστικές επιστήμες και το ιστορικό / πολιτιστικό περιεχόμενο (Πανδέκτης). Ιδιαίτερη έμφαση δόθηκε στη περιγραφή του περιεχομένου των αποθετηρίων καθώς και σε τεχνικά θέματα κυρίως όσον αφορά επεκτάσεις σε ΕΛ/ΛΑΚ πλατφόρμες αποθετηρίων και στην ανάπτυξη και συντήρηση της απαραίτητης υποδομής συστημάτων.

Δράσεις που σχεδιάζονται για το μέλλον περιλαμβάνουν μεταξύ άλλων την ακόμα μεγαλύτερη βελτίωση και αυτοματοποίηση των διαδικασιών αυτό-αρχειοθέτησης ώστε να γίνουν ελκυστικότερες για τους παραγωγούς περιεχομένου (π.χ. ερευνητές, ερευνητικούς οργανισμούς), την ανάπτυξη υπηρεσιών προστιθέμενης αξίας που θα αποτελέσουν ένα επιπλέον κίνητρο τακτικότερης χρήσης των αποθετηρίων από την ερευνητική κοινότητα καθώς και την περαιτέρω διερεύνηση και επίλυση θεμάτων σχετικών με πνευματικά δικαιώματα ώστε να εξασφαλιστεί η ανοικτή πρόσβαση σε όσο το δυνατόν περισσότερο επιστημονικό περιεχόμενο. Σε πιο τεχνικό επίπεδο βρίσκονται σε εξέλιξη ενέργειες για τη διάθεση των διάφορων εφαρμογών και λειτουργιών που έχουν αναπτυχθεί σε κάθε ενδιαφερόμενο. Αυτό επιτυγχάνεται μέσω της σταδιακής προσθήκης σχετικής τεκμηρίωσης στο wiki της ελληνικής κοινότητας χρηστών του DSpace που διατηρείται από τη Βιβλιοθήκη του Πανεπιστημίου Πατρών [18], ενώ για μεγαλύτερου εύρους επεκτάσεις εξετάζεται η αξιοποίησή τους ως ΕΛ/ΛΑΚ είτε αυτόνομα είτε ως τμήμα μεγαλύτερων συστημάτων (π.χ. του DSpace). Επιπλέον, το ΕΚΤ σκοπεύει να συνεχίσει με αυξανόμενη ένταση τις δράσεις για την προώθηση της ιδέας της ανοικτής πρόσβασης ώστε να εξασφαλιστεί η όσον το δυνατόν μεγαλύτερη αποδοχή και υιοθέτησή της από την ελληνική ερευνητική κοινότητα.

ΕΥΧΑΡΙΣΤΙΕΣ

Η ανάπτυξη των αποθετηρίων Ήλιος και Πανδέκτης πραγματοποιήθηκε στο πλαίσιο της «Ψηφιακής Ελλάδας» (έργα: «Εθνικό Πληροφοριακό Σύστημα Έρευνας και Τεχνολογίας (ΕΠΣΕ+Τ – Γ’ Φάση» και «Πανδέκτης: Ψηφιακός Θησαυρός Πρωτογενών Τεκμηρίων Ελληνικής Ιστορίας και Πολιτισμού») που συγχρηματοδοτήθηκαν από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης) και το Ελληνικό Δημόσιο (Επιχειρησιακό Πρόγραμμα «Κοινωνίας της Πληροφορίας», Γ’ ΚΠΣ). Οι συγγραφείς επιθυμούν να εκφράσουν τις ευχαριστίες τους στην διεπιστημονική ομάδα συνεργατών που συνέβαλλε στην υλοποίηση των δράσεων που περιγράφονται στο παρόν άρθρο.

ΑΝΑΦΟΡΕΣ

[1] Barton, M., M. Waters (2005). LEADIRS: Learning About Digital Institutional Repositories, Creating an Institutional Repository. MIT Libraries.
[2] Crow, R (2002). The case of Institutional repositories: A SPARC Position Paper.
[3] Lam K.-T., D.L.H. Chan (2007). Building an institutional repository: Sharing experiences at the HKUST library. OCLC Systems and Services, 23 (3), pp. 310-323.
[4] Greig, M. and Nixon, W.J. (2007) On the road to Enlightenment: establishing an institutional repository service for the University of Glasgow. OCLC Systems & Services 23(3):pp. 297-309.
[5] Salo, D. (2009, April). Name authority control in institutional repositories. Cataloging and Classification Quarterly 47 (3/4).
[6] Phillips, S., J. Creel, C. Green, Y. Li, A. Maslov, P. Mattingly, A. Mikeal, J. Paz, J. Leggett, and M. McFarland (2008, April). Manakin: Lessons learned. In Third International Conference on Open Repositories, Southampton, UK.
[7] Saidis, K., G. Pyrounakis, M. Nikolaidou, and A. Delis (2006). Digital object prototypes: An effective realization of digital object types. pp. 123–134.
[8] Πυρουνάκης, Γ., Κ. Σαΐδης, Κ. Βίγλας, Ε. Λουρδή, και Μ. Νικολαΐδου (2006, Νοέμβριος). Αναπαράσταση και Διαχείριση Ετερογενών Ψηφιακών Συλλογών στο Σύστημα Ψηφιακής Βιβλιοθήκης Πέργαμος. Πρακτικά 15ου Πανελλήνιου Συνέδριου Ακαδημαϊκών Βιβλιοθηκών.

[9] Huynh, D. F., D. R. Karger, and R. C. Miller (2007). Exhibit: lightweight structured data publishing. In WWW '07: Proceedings of the 16th international conference on World Wide Web, New York, NY, USA, pp. 737-746. ACM.
[10] Allinson, J., S. François, and S. Lewis (2008, January). SWORD: Simple web-service offering repository deposit. *Ariadne* (54).
[11] Office of Government Commerce, ICT Infrastructure Management. The Stationery Office, 2002.
[12] ISO/IEC 27001:2005 committee, “Information technology -- Security techniques -- Specification for an Information Security Management System”, 2005.
[13] Rosenblum, M. and T. Garfinkel (2005) “Virtual machine monitors: Current Technology and Future Trends”, IEEE Internet Computing, May 2005, Vol. 38, No. 5.
[14] Koomey, J.G. “Estimating Total Power Consumption by Servers in the U.S. and the World”, Technical Report Final Report, Lawrence Berkeley National Laboratory, February 2007, http://hightech.lbl.gov/documents/DATA_CENTERS/svrpwrusecompletefinal.pdf
[15] Stathopoulos, P., A. Soumplis, N. Houssos (2009). The case study of an F/OSS virtualization platform deployment and quantitative results, Open Source Ecosystems: Diverse Communities Interacting, 5th IFIP WG 2.13 International Conference on Open Source Systems, OSS 2009, Skövde, Sweden, June 3-6, 2009. Proceedings, IFIP 299 Springer 2009.
[16] MADS specification, http://www.loc.gov/standards/mads/
[17] Authority Control of Metadata Values. DSpace Wiki, http://wiki.dspace.org/index.php/Authority_Control_of_Metadata_Values
[18] Wiki της Ελληνικής Κοινότητας χρηστών του DSpace, Βιβλιοθήκη & Κέντρο Πληροφόρησης - Πανεπιστήμιο Πατρών, http://nemertes.lis.upatras.gr/wiki/.