Αρχειοθέτηση του Ελληνικού Ιστού

Δήμητρα Χιώτη[1], Μιχάλης Βαζιργιάννης[2], Πολύκαρπος Μελαδιανός[3], Γεωργία Αγγελάκη[4]

Περίληψη

Στόχος του έργου είναι να συγκεντρώσει το μέγιστο δυνατό υποσύνολο του Ιστού στην Ελληνική γλώσσα, λειτουργώντας ως αρχειοθετικός μηχανισμός του ελληνικού Διαδικτύου, με απώτερο σκοπό να προάγει μακροπρόθεσμα την επιστημονική έρευνα. Για την επίτευξη του παραπάνω στόχου έχει δημιουργηθεί η διεπαφή χρήστη/ βιβλιοθηκονόμου του Εθνικού Συστήματος Αρχειοθέτησης του Ελληνικού Ιστού (ΕΣΑΕΙ). Μέσω της διεπαφής, τα μέλη της Ομάδας Αρχειοθέτησης Ιστού έχουν τη δυνατότητα να αναζητήσουν ιστοσελίδες με λέξεις κλειδιά, θεματική κατηγορία ή όνομα ιστοχώρου (URL) για την ανάκτηση και ανασύσταση ιστοσελίδων όπως ήταν στο παρελθόν - μερικές δεν είναι πια διαθέσιμες online - και να διαπιστώσουν την εξέλιξη των ελληνικών ιστοτόπων. Προς το παρόν, δεν διατίθεται πρόσβαση στο κοινό.

Μέχρι στιγμής έχουν πραγματοποιηθεί δύο μαζικές συγκομιδές κειμένου του ελληνικού Ιστού οι οποίες ανά συγκομιδή έχουν συλλέξει κείμενο μεγέθους περίπου 14 τρισεκατομμυρίων χαρακτήρων και 7 εκατομμυρίων μοναδικών λέξεων, από περίπου 170 εκατομμύρια URIs ανά συγκομιδή. Εκτός της μαζικής συγκομιδής, που αφορά σε όλο το εθνικό domain (.gr), μέσω του ΕΣΑΕΙ παρέχεται η δυνατότητα αρχειοθέτησης ιστοσελίδων με βάση κάποιο θέμα ή γεγονός. Σχετικά, έχουν πραγματοποιηθεί επιλεκτικές συγκομιδές για τις θεματικές κατηγορίες «Τοπική Αυτοδιοίκηση», «Ειδησεογραφία» και «Παιδεία».

Λέξεις – Κλειδιά: Αρχειοθέτηση Ιστού, Εθνική Βιβλιοθήκη της Ελλάδος, Οικονομικό Πανεπιστήμιο Αθηνών, Ιστοσυγκομιδή, Ελληνικό domain (.gr)

---

[1] Βιβλιοθηκονόμος, MA, Εθνική Βιβλιοθήκη της Ελλάδος
[2] Καθηγητής, Οικονομικό Πανεπιστήμιο Αθηνών & École Polytechnique
[3] PhD, Οικονομικό Πανεπιστήμιο Αθηνών
[4] Υπεύθυνη Ανάπτυξης Ψηφιακών Υπηρεσιών της Εθνικής Βιβλιοθήκης της Ελλάδος, στο πλαίσιο του «Προγράμματος Μετεγκατάστασης της ΕΒΕ στο ΚΠΙΣΝ 2015-2018»

1. Εισαγωγή

Το Διαδίκτυο αποτελεί μοναδική πηγή πληροφοριών, αφού καταγράφει τη συλλογική και ατομική δράση των πολιτών και τον τρόπο που αλληλεπιδρούν οι μεμονωμένοι χρήστες και οι κοινότητες χρηστών μεταξύ τους, χρησιμοποιώντας εξειδικευμένη τεχνολογία. Ο Μπάνος (2015) αναφέρει ότι η μέση διάρκεια ζωής των ιστοσελίδων είναι μικρότερη από 100 ημέρες. Δεδομένης της ραγδαίας αύξησης του όγκου του διαδικτύου και του πεπερασμένου χρόνου ζωής των ιστοσελίδων, η θεσμοθέτηση της Αρχειοθέτησης του Ιστού για τους σκοπούς της διάσωσης της πολιτισμικής κληρονομιάς και της συλλογικής μνήμης και η χρησιμοποίησή της για τους σκοπούς της έρευνας και της ανάπτυξης, αποτελεί, εδώ και χρόνια, αναγκαιότητα για πολλές Εθνικές Βιβλιοθήκες.

Στην Ελλάδα, η πρώτη προσπάθεια Αρχειοθέτησης του Ιστού αφορούσε τις ιστοσελίδες του Οικονομικού Πανεπιστημίου Αθηνών και ξεκίνησε το 2010[5] (Plachouras, 2010). Η Αρχειοθέτηση του Ιστού θεσμοθετήθηκε επίσημα με τον ν. 4452/2017 «Περί ρύθμισης θεμάτων Κρατικού Πιστοποιητικού Γλωσσομάθειας, θεμάτων της ΕΒΕ και άλλες διατάξεις», ο οποίος τροποποίησε το άρ.1(2-4) του ν.3149/2002, μια διάταξη με γενική διατύπωση για την ιστοσυγκομιδή ως αρμοδιότητα της Εθνικής Βιβλιοθήκης της Ελλάδος (ΕΒΕ). Σύμφωνα με τον νόμο 4452/2017[6], η ΕΒΕ:

> «Λειτουργεί ως το επίσημο Εθνικό Αποθετήριο και Αρχείο ψηφιακών δημοσιευμάτων, δεδομένων και μεταδεδομένων που παράγονται στη χώρα ή αφορούν τον ελληνικό πολιτισμό. Στην ανωτέρω λειτουργία περιλαμβάνεται η παρακολούθηση και η αρχειοθέτηση του παγκόσμιου ιστού (web archiving) ή άλλου τεχνολογικού περιβάλλοντος. Προς τούτο αναλαμβάνει, κατανέμει και συντονίζει σε εθνικό επίπεδο τις σχετικές δράσεις.»

Η ανάπτυξη και η πιλοτική λειτουργία της Υπηρεσίας χρηματοδοτήθηκε με αποκλειστική χορηγία από το Ίδρυμα Σταύρος Νιάρχος στο πλαίσιο του Προγράμματος "Σχέδιο Υλοποίησης Μετεγκατάστασης της ΕΒΕ στο Κέντρο Πολιτισμού Ίδρυμα Σταύρος Νιάρχος 2015-2017". Είναι ένα από τα 17 έργα της Δράσης 2 που αφορούσε στην ανάπτυξη των Ψηφιακών Υπηρεσιών της

[5] Αρχειοθέτηση Ιστοσελίδων Οικονομικού Πανεπιστημίου Αθηνών. http://archive.aueb.gr/

[6] ν. 4452/2017. https://www.e-nomothesia.gr/kat-ekpaideuse/nomos-4452-2017-phek-17a15-2-2017.html

Εθνικής Βιβλιοθήκης, και ένα από 50 έργα που ολοκληρώθηκαν στο πλαίσιο υλοποίησης του Προγράμματος.

Το έργο πραγματοποιείται μέσω της συνεργασίας της Ομάδας Αρχειοθέτησης Ιστού της ΕΒΕ με την ερευνητική ομάδα "Εξόρυξης Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό" του Οικονομικού Πανεπιστημίου Αθηνών (ΟΠΑ).

2. Προκλήσεις υλοποίησης έργου

Κατά την διαδικασία υλοποίησης του έργου, οι εμπλεκόμενοι φορείς κλήθηκαν να αντιμετωπίσουν διάφορες προκλήσεις όπως ο καθορισμός του όγκου του ελληνικού ιστού, τα ιδιαίτερα χαρακτηριστικά του, η εύρεση του απαραίτητου υλισμικού και λογισμικού (π.χ. αποθηκευτικοί χώροι, προγράμματα) για την συγκομιδή, επεξεργασία, αποθήκευση, εμφάνιση και διατήρηση του Αρχείου Ιστού, το βάθος και η συχνότητα συγκομιδής και οι δυνατότητες αναζήτησης και χρήσης. Ακόμη, διερευνήθηκαν νομικά ζητήματα, όπως τα πνευματικά δικαιώματα των ιστοτόπων και του περιεχομένου τους, τα προσωπικά δεδομένα που εξάγονται ως πληροφορία από τον Ελληνικό Ιστό και τα δικαιώματα πρόσβασης και χρήσης του Αρχείου Ιστού.

3. Φάσεις υλοποίησης έργου

3.1. Πρώτη φάση

Στο ξεκίνημα του έργου πραγματοποιήθηκε οικονομοτεχνική μελέτη με στόχο την ανάπτυξη ενός Εθνικού Συστήματος Αρχειοθέτησης του Ιστού. Για τις ανάγκες της μελέτης έγινε συλλογή ενός αντιγράφου του συνόλου του ελληνικού Ιστού, μόνο ως προς το κείμενο. Ως “ελληνικοί” θεωρήθηκαν οι ιστότοποι που:

- Ανήκαν στο domain .gr και ήταν γραμμένοι στην ελληνική ή/και άλλη γλώσσα
- Δεν ανήκαν στο domain .gr και ήταν γραμμένοι στα ελληνικά (π.χ. ανήκαν στα domains .edu ή .com)

Δεν συλλέχθηκαν ιστότοποι που:

- Ακολουθούσαν το πρωτόκολλο Robots Exclusion Protocol[7] (περιελάμβαναν αρχεία .txt)
- Αποτελούσαν πολυμεσικό πόρο (εικόνα, βίντεο, ήχο)

[7] Robots Exclusion Protocol. http://www.robotstxt.org/orig.html

- Είχαν τοποθετηθεί εκτός του πεδίου ορισμού, δηλαδή δεν ανήκαν στο domain .gr ούτε ήταν γραμμένοι στα ελληνικά
- Απαιτούσαν κωδικούς πρόσβασης

Οι πρωταρχικοί σύνδεσμοι που χρησιμοποιήθηκαν για την πρώτη συγκομιδή ήταν γύρω στις 45.000 και προέκυψαν από έρευνα σε μηχανές αναζήτησης και σχετικές με τον ελληνικό ιστό θεματικές πύλες. Αυτοί χρησιμοποιήθηκαν για την εύρεση του συνόλου του ελληνικού διαδικτύου. Παράλληλα με αυτές τις εργασίες, διερευνήθηκαν οι διεθνείς καλές πρακτικές ως προς τα εργαλεία και τις μεθόδους της Ιστοσυγκομιδής.

Η συλλογή του πρώτου "αποτυπώματος" του ελληνικού Ιστού διήρκησε 27 ημέρες και 9 ώρες. Η συγκομιδή πραγματοποιήθηκε σε βάθος πέντε (5), κάτι που επέτρεψε τη διαστασιολόγηση του Ελληνικού Ιστού. Προσπελάστηκαν συνολικά 128.992 πρωταρχικοί σύνδεσμοι που αντιστοιχούσαν σε 247.910.539 επεξεργασμένα URIs, εκ των οποίων η επιτυχημένη συλλογή αφορούσε τα 232.203.535 URIs και αντιστοιχούσε σε έναν όγκο συλλογής 18TB, που κατόπιν συμπίεσης ισούται με 3,2 TB. Επιπλέον, έγιναν αναλύσεις του περιεχομένου και εντοπίστηκαν τα ηλεκτρονικά καταστήματα, τα οποία αντιστοιχούσαν στο 45,3% του συνολικού όγκου. Αυτά αφαιρέθηκαν από τη συλλογή του ιστοπεριεχομένου.

Από την πρώτη μαζική ιστοσυγκομιδή προέκυψε η δημιουργία του γράφου του ελληνικού διαδικτύου, που αποτυπώνει με οπτικό τρόπο την τοπολογία της διασύνδεσης μεταξύ σημαντικών ιστοτόπων (ως προς τον αριθμό των εισερχόμενων και εξερχόμενων συνδέσμων), αναδεικνύοντας θεματικές και άλλου τύπου συσχετίσεις. Επιπλέον, τέτοιου είδους γράφοι θα μπορούσαν να χρησιμοποιηθούν μελλοντικά για την οπτικοποίηση της συλλογής από τους χρήστες. Η μελέτη κατέδειξε επίσης, ότι απαιτούνται τουλάχιστον 450 TB αποθηκευτικού χώρου, συμπεριλαμβανομένων των αντιγράφων ασφαλείας, σε βάθος 5ετίας, για την ανάπτυξη μιας ολοκληρωμένης Εθνικής υποδομής Ιστοσυγκομιδής.

3.2. Δεύτερη φάση

Η δεύτερη φάση ανάπτυξης εστίασε στην εγκατάσταση, παραμετροποίηση και λειτουργία σε παραγωγική μορφή των βασικών εργαλείων του Εθνικού Συστήματος Αρχειοθέτησης του Ιστού, που προέκυψαν από τη μελέτη, σε υποδομές της Εθνικής Βιβλιοθήκης. Στόχος του έργου ήταν να μπορεί η ΕΒΕ να εξυπηρετήσει τις ανάγκες μαζικής και στοχευμένης συγκομιδής του Ελληνικού Ιστού, τις απαιτήσεις επιμέλειας του περιεχομένου ως προς την ταξινόμηση και οργάνωσή του από τους βιβλιοθηκονόμους της ΕΒΕ, και τις ανάγκες αναζήτησης και πλοήγησης στο περιεχόμενο για το χρήστη.

Η ανάπτυξη της αρχιτεκτονικής του Συστήματος βασίστηκε σε διαδεδομένες τεχνολογίες ανοικτού λογισμικού, όπως Heritrix για τη συγκομιδή, Solr για την ευρετηρίαση και Open Wayback για την ανασύσταση των ιστοσελίδων, που χρησιμοποιούνται και από άλλες Εθνικές Βιβλιοθήκες. Με βάση τις παραπάνω τεχνολογίες και με τη χρήση της ολοκληρωμένης πλατφόρμας Netarchive Suite, αναπτύχθηκε μια πρωτότυπη διεπαφή για το χρήστη/βιβλιοθηκονόμο, στα ελληνικά. Η διεπαφή αυτή μετονομάστηκε σε Εθνικό Σύστημα Αρχειοθέτησης του Ελληνικού Ιστού (ΕΣΑΕΙ) και επιτρέπει τον προγραμματισμό και τη διενέργεια στοχευμένων και μαζικών συγκομιδών του Ιστού, την οργάνωση του περιεχομένου με βάση λέξεις-κλειδιά, την παρακολούθηση της κατάστασης των διαδικασιών ιστοσυγκομιδής και την εξαγωγή στατιστικών στοιχείων και αναφορών για την ιστοσυγκομιδή.

Παράλληλα, πραγματοποιήθηκε η δεύτερη συγκομιδή του ελληνικού ιστού (κείμενο μόνο). Για την δεύτερη συγκομιδή ακολουθήθηκαν οι ίδιες παράμετροι που ακολουθήθηκαν και για την πρώτη, με τη διαφορά ότι:

- Δόθηκε προτεραιότητα σε ιστότοπους που ανήκαν στο .gr
- Δεν συλλέχθηκαν ιστότοποι που θεωρήθηκαν ηλεκτρονικά καταστήματα.

4. Θεματικές Κατηγορίες

Στη δεύτερη φάση ανάπτυξης πραγματοποιήθηκαν οι πρώτες θεματικές συγκομιδές υλικού (με εικόνες), οι οποίες ανήκουν σε τρεις επιμέρους κατηγορίες: «Παιδεία», «Τοπική Αυτοδιοίκηση» και «Ειδησεογραφία». Οι θεματικές κατηγορίες αυτές επιλέχθηκαν με βάση τα παρακάτω κριτήρια:

Αναφορικά με την κατηγορία **Τοπική Αυτοδιοίκηση**, θεωρήθηκε πως θα έπρεπε να συμπεριληφθεί στις κατηγορίες που θα δημιουργηθούν σε πρώτη φάση, τόσο για τον έλεγχο της λειτουργίας του συστήματος, όσο και για τη μεγάλη σημασία της ανάδειξης της λειτουργίας της τοπικής αυτοδιοίκησης μέσα από την προβολή ιστοσελίδων περιεχομένου που αφορά στο δημόσιο συμφέρον. Η ομάδα έργου βασίστηκε σε έτοιμη λίστα 350 πρωταρχικών διευθύνσεων, που συγκεντρώθηκαν χειρωνακτικά, με όλους τους ιστοτόπους των Περιφερειών, των Δήμων, των Αποκεντρωμένων Διοικήσεων και των Κοινοτήτων του Δήμου Αθηναίων.

Η κατηγορία **Ειδησεογραφία** θεωρήθηκε πολύ χρήσιμη εξαιτίας της συχνής ανανέωσης του περιεχομένου που εμπεριέχουν οι συγκεκριμένοι ιστότοποι και του πλούτου των διαθέσιμων πληροφοριών που παρέχει. Μέσω της ιστοσελίδας

της Γενικής Γραμματείας Ενημέρωσης και Επικοινωνίας[8] συγκεντρώθηκε μια λίστα με ειδησεογραφικούς ιστοτόπους. Από αυτούς αφαιρέθηκαν τα σφάλματα και όσες σελίδες κρίθηκε πως δεν ήταν ειδησεογραφικές και σε αυτές που έμειναν προστέθηκε τίτλος και υποκατηγορία (-ες). Συνολικά συλλέχθηκαν 661 ιστότοποι, ορισμένοι από τους οποίους αποτελούν ψευδότιτλο.

Η κρίση για το ποια ιστοσελίδα θεωρείται ειδησεογραφική και ποια όχι στηρίχτηκε στους ορισμούς για τις έννοιες "είδηση" και "ειδησεογραφία" που δίνονται από έγκυρα λεξικά[9][10]. Οι υποκατηγορίες επιλέχθηκαν πολύ προσεκτικά, με βάση τα θέματα που αναφέρονταν στον κάθε ιστότοπο. Οι ιστότοποι με αντίστοιχο έντυπο που διαθέτει ή δεν διαθέτει η ΕΒΕ, καταχωρίστηκαν σε ξεχωριστό αρχείο .xls, κάτι που μπορεί να αποτελέσει προεργασία για μελλοντικό εμπλουτισμό της τεκμηρίωσης της συλλογής στη Διεπαφή Βιβλιοθηκονόμου.

Η θεματική κατηγορία **Παιδεία** ορίστηκε με βάση το περιεχόμενο της πρώτης και δεύτερης συγκομιδής του ελληνικού ιστού. Αποτελείται από σελίδες που ανήκουν στο .gr και αφορούν στην παιδεία (ιστοσελίδες σχολείων, πανεπιστημίων, φροντιστηρίων κλπ.). Σε πρώτη φάση προστέθηκαν όσοι ιστότοποι ανήκουν στο edu.gr, στο sch.gr και στο mysch.gr. Αυτές οι σελίδες αποτελούν και τη μεγάλη πλειοψηφία των ιστοτόπων της θεματικής κατηγορίας. Στη συνέχεια προστέθηκαν κάποιοι ιστότοποι που ανήκουν στην κατηγορία παιδεία, με βάση την ταξινόμηση που έγινε στα πλαίσια της ανάλυσης των δεδομένων της πρώτης συγκομιδής. Συνολικά, η συλλογή αυτή αποτελείται από 8.199 ιστοτόπους.

Μέσω αυτής της κατηγορίας θα παρέχεται στους χρήστες η δυνατότητα πλοήγησης και μελέτης της εξέλιξης αντιπροσωπευτικού μέρους ιστοσελίδων της ελληνικής παιδείας και του ελληνικού πολιτισμού. Η συγκεκριμένη κατηγορία, θα μπορούσε να χρησιμοποιηθεί σε επόμενη φάση του έργου για περαιτέρω υποκατηγοριοποίηση και συστήνεται για εκπαίδευση του συστήματος και πειραματισμό, με σκοπό τον εμπλουτισμό της συλλογής.

---

[8] Γενική Γραμματεία Ενημέρωσης και Επικοινωνίας. Μητρώο Online Media. https://emedia.media.gov.gr/

[9] Στο Λεξικό της Κοινής Νεοελληνικής του ιστοτόπου http://www.qreek-lanquage.qr δίνεται ο παρακάτω ορισμός για τον όρο "Ειδησεογραφία"'
ειδησεογραφία η 025 .• α.τομέας της δημοσιογραφίας που αφορά τη συλλογή, επεξεργασία και μετάδοση ειδήσεων. β. το σύνολο των ειδήσεων που αναγράφονται σε εφημερίδα ή περιοδικό, ή που μεταδίδονται από το ραδιόφωνο και την τηλεόραση: Αντικειμενική / πλούσια / πολιτική / καλλιτεχνική -e . Σελίδες ειδησεογραφίας.

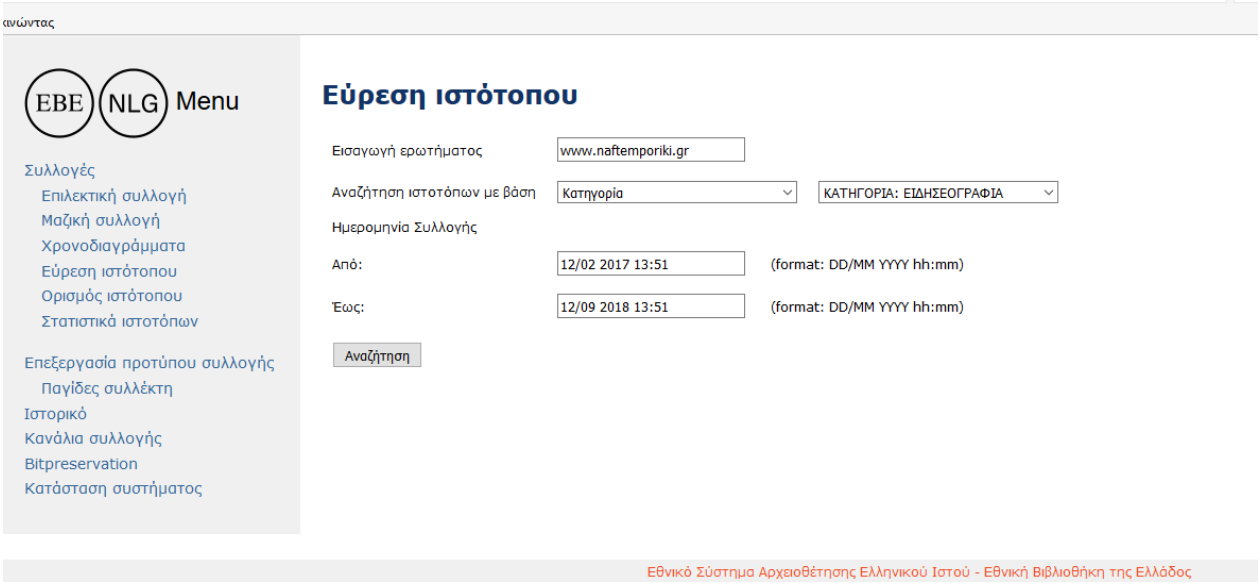
[10] Στο Λεξικό Τριανταφυλλίδη δίνεται ο παρακάτω ορισμός για τον όρο "Είδηση":
είδηση η [ίδίςί]: 1.λόγος που γνωστοποιεί πρόσφατο γεγονός ή συμβάν, πληροφορεί

5. Διεπαφή ΕΣΑΕΙ - Λειτουργία εύρεσης ιστοτόπου

Η αναζήτηση στο περιεχόμενο του Ελληνικού Αρχείου Ιστού πραγματοποιείται μέσω της διεπαφής χρήστη/ βιβλιοθηκονόμου του ΕΣΑΕΙ και αφορά μόνο στην επιλεκτική συγκομιδή που έχει πραγματοποιηθεί, δηλαδή στις θεματικές κατηγορίες. Οι δυνατότητες αναζήτησης που παρέχονται είναι με:

- Θεματική Κατηγορία
- Λέξεις – Κλειδιά
- Url

Υπάρχει επίσης η δυνατότητα περιορισμού των αποτελεσμάτων σε συγκεκριμένα χρονικά διαστήματα.



Εικόνα 1. Η λειτουργία εύρεσης ιστοτόπου του ΕΣΑΕΙ (αναζήτηση με Κατηγορία).

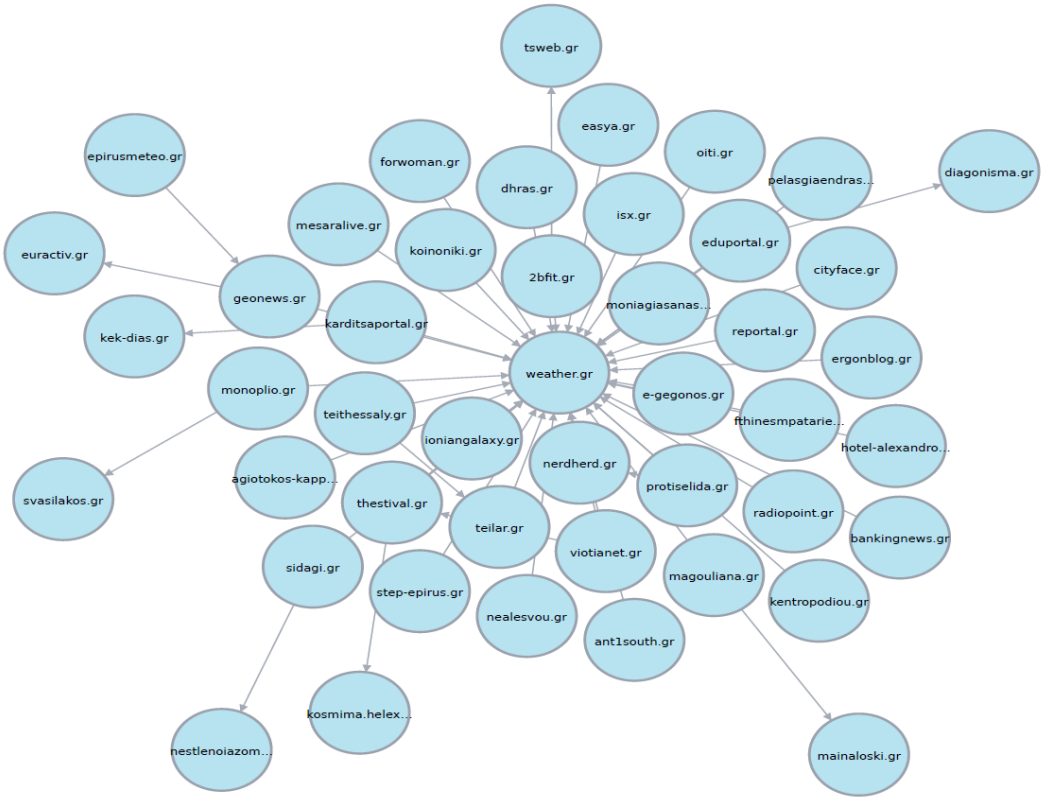
Ήδη με τις πρώτες προσπάθειες συγκομιδής του ελληνικού ιστού, έχει δημιουργηθεί, πιθανότατα, το μεγαλύτερο σώμα κειμένων στην ελληνική γλώσσα που υπήρξε ποτέ, συγκεντρωμένο σε ψηφιακή μορφή προς επεξεργασία και εξαγωγή γνώσης. Υπολογίζεται πως ανά συγκομιδή έχει συλλεχθεί κείμενο μεγέθους περίπου 14 τρισεκατομμυρίων χαρακτήρων και 7 εκατομμυρίων μοναδικών λέξεων, από περίπου 170 εκατομμύρια URIs ανα συγκομιδή. Το υλικό αυτό παρουσιάζει τεράστιο ενδιαφέρον για την έρευνα, σε συνδυασμό με τις νέες δυνατότητες επεξεργασίας που επιτρέπουν οι τεχνολογίες της τεχνητής νοημοσύνης, της εξόρυξης δεδομένων και της μηχανικής μάθησης, σε πολλούς τομείς όπως γλωσσολογία, ιστορία, κοινωνιολογία, κ.λπ., καθώς και για την παραγωγή υπηρεσιών προστιθέμενης αξίας.

Εντούτοις, αν και η ΕΒΕ ανέλαβε την υποχρέωση της ιστοσυγκομιδής, το ισχύον πλαίσιο για την προστασία της πνευματικής ιδιοκτησίας, δεν επιτρέπει προς το παρόν, τη διάθεση του συλλεγόμενου υλικού, όπως κατέδειξε η σχετική έρευνα που πραγματοποιήθηκε από το Νομικό Σύμβουλο της ΕΒΕ (Παπαδόπουλος, 2017). Για το λόγο αυτό, η ιστοσυγκομιδή στην Ελλάδα πραγματοποιείται με τον σκοπό της διαφύλαξης της ελληνικής πολιτισμικής κληρονομιάς του διαδικτύου, ώστε να είναι αυτή διαθέσιμη στον μελλοντικό ερευνητή.

6. Ανάπτυξη συστήματος

Το Εθνικό Σύστημα της Αρχειοθέτησης του Ελληνικού Ιστού παρουσιάζει τεράστιο ενδιαφέρον για περαιτέρω ανάπτυξη, εφόσον εξασφαλιστεί κατάλληλη χρηματοδότηση. Κύριο ερευνητικό ενδιαφέρον παρουσιάζουν τα παρακάτω θέματα:

- η κατηγοριοποίηση και θεματική ταξινόμηση ιστοτόπων με βάση το περιεχόμενό τους σε συγκεκριμένες κατηγορίες πολιτιστικές, ενημερωτικές, blog, με βάση τα ατομικά και συλλογικά ενδιαφέροντα των χρηστών.
- η περαιτέρω ανάπτυξη των υπηρεσιών προστιθέμενης αξίας με τη διάθεση προγραμματιστικών διεπαφών για την πρόσβαση στο ιστοπεριεχόμενο και το γράφημα, που θα επιτρέπουν, π.χ. τον εντοπισμό της συχνότητας εμφάνισης επιμέρους λέξεων και τις συσχετίσεις μεταξύ ιστοτόπων.
- η δημιουργία μιας πλήρους εργαλειοθήκης αναφοράς της ελληνικής γλώσσας - γλώσσα με παγκόσμια σημασία - όπως η παραγωγή ενός πλήρους λεξικού της ελληνικής γλώσσας, η δημιουργία μιας λίστας με περιττές λέξεις (stopwords), προγράμματος που εξάγει τις ρίζες των λέξεων (stemmer), ο εντοπισμός θέσεων των ελληνικών λέξεων σε έναν διανυσματικό χώρο πιο γνωστό στα αγγλικά σαν word embeddings, κ.λ.π., εργαλεία βασικά για περαιτέρω εφαρμογές στην έρευνα και την υπολογιστική ευφυΐα.



Εικόνα 2. Γράφημα που απεικονίζει δείγμα της τοπολογίας του ελληνικού διαδικτύου, με επίκεντρο τον ιστότοπο weather.gr.

7. Χρήστες και πεδία ενδιαφέροντος

Παραδείγματα χρηστών/ ομάδων χρηστών που θα μπορούσαν να έχουν εξειδικευμένο ενδιαφέρον για το Αρχείο Ιστού είναι τα παρακάτω:

- Ιστορικοί. Μάλιστα ένας νέος και πολλά υποσχόμενος κλάδος της ιστορίας είναι η λεγόμενη «Web Historiography» (Brügger, 2012)
- Κοινωνιολόγοι
- Γλωσσολόγοι
- Δημοσιογράφοι
- Κάτοχοι/ σχεδιαστές ιστοσελίδων
- Δημόσια ιδρύματα που δεν αρχειοθετούν τις ιστοσελίδες τους και παραπέμπουν το κοινό στο διαδίκτυο
- Ερευνητές των κοινωνικών επιστημών που ασχολούνται με το διαδίκτυο.

Σύμφωνα με σχετική έρευνα (Dougherty, 2010) για τη χρήση των Αρχείων Ιστού, τα πιθανά πεδία ενδιαφέροντος μπορεί να είναι το διαδίκτυο συνολικά, οι

κοινότητες που προκύπτουν από τις συνδέσεις των σελίδων, ένας ιστότοπος, μια ιστοσελίδα ή ένα στοιχείο μέσα στην ιστοσελίδα.

8. Συμμετοχή στο IIPC

Από τον Ιούλιο του 2018, η Ομάδα Αρχειοθέτησης Ιστού της ΕΒΕ συμμετέχει ενεργά στο International Internet Preservation Consortium (IIPC), ένα διεθνή οργανισμό που προωθεί την συνεργασία σε παγκόσμιο επίπεδο για τα θέματα της αρχειοθέτησης ιστού. Με αυτό τον τρόπο έχει άμεση πρόσβαση σε μια πληθώρα πληροφοριών, όπως τα χρησιμοποιούμενα εργαλεία, η εξέλιξή τους, τα χρησιμοποιούμενα πρότυπα και οι διεθνείς πολιτικές και καλές πρακτικές ως προς την συγκομιδή, αποθήκευση, επεξεργασία, διατήρηση, αναζήτηση και παρουσίαση των αποτελεσμάτων. Επιπλέον, έχει τη δυνατότητα να συμμετέχει σε προγράμματα συγκομιδών διεθνούς εμβέλειας, όπως το "The Online News around the World Project" και σε διεθνείς ομάδες εργασίας με σκοπό την βελτίωση των μέσων και των όρων της αρχειοθέτησης. Εκτός αυτού, η Ελλάδα γνωστοποιεί με αυτό τον τρόπο την παρουσία της στη διεθνή κοινότητα Αρχειοθέτησης Ιστού και αναλαμβάνει ενεργό ρόλο στην εξέλιξη των διεθνών πολιτικών και πρακτικών.

9. Επόμενα σχέδια

Στα επόμενα σχέδια της Ομάδας Αρχειοθέτησης Ιστού της ΕΒΕ εντάσσονται μεταξύ άλλων:

→ Η μαζική συγκομιδή του ελληνικού domain (.gr και .ελ) με πολυμέσα καθώς και η δημιουργία επιπλέον συλλογών - θεματικών κατηγοριών.

→ Η περεταίρω ανάπτυξη της διεπαφής χρήστη/ βιβλιοθηκονόμου με εξαγωγή επιπλέον στατιστικών, εισαγωγή εργαλείου αυτόματου ποιοτικού ελέγχου, μετάφραση της πλατφόρμας κλπ.

→ Η συνεργασία με εθνικούς φορείς, τόσο για την εξεύρεση χρηματοδότησης για την συνέχιση του έργου από εθνικούς, ευρωπαϊκούς και ιδιωτικούς πόρους), όσο και για την πραγματοποίηση του έργου της Αρχειοθέτησης Ιστού σε πρακτικό επίπεδο (π.χ. συνεργασία με την Εθνική Επιτροπή Τηλεπικοινωνιών και Ταχυδρομείων (ΕΕΤΤ) για την εύρεση των πρωταρχικών συνδέσμων για την συγκομιδή του ελληνικού domain (πλέον .gr και .ελ).

→ Η συνεργασία με διεθνείς οργανισμούς αρχειοθέτησης ιστού, όπως το Internet Archive, για ζητήματα όπως ο εμπλουτισμός θεματικών κατηγοριών με σελίδες «ελληνικού περιεχομένου» που δημοσιεύονται στο εξωτερικό. Μάλιστα, η επιλογή της συγκεκριμένης κατηγορίας συνάδει με τον Ν. 4452/2017, σύμφωνα με τον οποίο θα πρέπει να συλλέγονται: «...τεκμήρια

που δημιουργούνται ή δημοσιεύονται στο εξωτερικό αλλά συνδέονται ή έχουν συνάφεια με τους ανθρώπους, τη χώρα, τις παραδόσεις και τον πολιτισμό της Ελλάδας».
→ Η ενεργός συμμετοχή στις ομάδες έργου του IIPC, με σκοπό την βελτιστοποίηση των πρακτικών αρχειοθέτησης για το Ελληνικό Αρχείο Ιστού.
→ Η δημιουργία ψηφιακών γλωσσολογικών πόρων μεγάλης κλίμακας, απαραίτητων για την έρευνα στην υπολογιστική γλωσσολογία αλλά και την βιομηχανική/οικονομική δραστηριότητα που σχετίζεται με γλώσσα και κείμενο. Οι πόροι αυτοί περιλαμβάνουν:

1. Λεξικό με τις μοναδικές λέξεις που βρέθηκαν (υπάρχει μια πρώτη έκδοση περίπου 7 εκατομμυρίων λέξεων)
2. Ενθέσεις λέξεων (word embeddings), απεικονίσεις σε διανυσματικούς χώρους - με χρήση βαθέων νευρωνικών δικτύων - όπου αναδεικνύεται η σημασιολογική ομοιότητα λέξεων με βάση την συχνή συνύπαρξη τους σε κείμενα (naftemporiki.gr, 2018).

10. Συμπεράσματα

Η Αρχειοθέτηση του Ιστού αποτελεί κοινή πρακτική δεκάδων χωρών ανά τον κόσμο εδώ και δεκαετίες για την καταγραφή της πολιτισμικής κληρονομιάς του Διαδικτύου. Ο νόμος 4452/2017 επισημοποιεί ουσιαστικά την αναγκαιότητα της καταγραφής αυτής για την Ελλάδα.

Η ΕΒΕ, ακολουθώντας τον θεσμικό ρόλο της, έχει ξεκινήσει ήδη την συγκομιδή του ελληνικού domain .gr και την πραγματοποίηση θεματικών συγκομιδών, για την κάλυψη πιο εξειδικευμένων αναγκών. Επίσης, μέσω συνεργασιών σε εθνικό και διεθνές επίπεδο, στοχεύει στην βελτιστοποίηση των μεθόδων Αρχειοθέτησης και στην περεταίρω αξιοποίηση του παραγόμενου περιεχομένου. Έτσι, μέσα από όλες τις ενέργειές του, το Ελληνικό Αρχείο Ιστού φιλοδοξεί να αποτελέσει τον θεματοφύλακα της γνώσης που παράγεται μέσω του ελληνικού Διαδικτύου και σημείο αναφοράς για την επιστημονική κοινότητα του μέλλοντος.

Βιβλιογραφία

1. Μπάνος, Β. (2015). *Web crawling, analysis and archiving. Diss*. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ). Σχολή Θετικών Επιστημών. Τμήμα Πληροφορικής. Εργαστήριο Τεχνολογίας και Επεξεργασίας Δεδομένων,. Ανάκτηση από http://ikee.lib.auth.gr/record/276644/files

2. Παπαδόπουλος, Μ. (2017). Γνωμοδότηση επί των νομικών πτυχών της ανάπτυξης και διάθεσης Εθνικής Υπηρεσίας Αρχειοθέτησης του Ελληνικού. Αθήνα.
3. Brügger, N. (2012). "When the present web is later the past: Web historiography, digital history, and internet studies.". *Historical Social Research/Historische Sozialforschung*, σσ. 102-117. Ανάκτηση από https://www.jstor.org/stable/41756477?seq=1#metadata_info_tab_contents
4. Dougherty, M. (2010). *Researcher engagement with web archives: State of the art*. Ανάκτηση από https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997
5. naftemporiki.*gr*. (2018, Μαϊ 07). Ανάκτηση από ΟΠΑ: Η τεχνητή νοημοσύνη στην υπηρεσία γλωσσολογικών πόρων για την Ελληνική: https://m.naftemporiki.gr/story/1347787/protopora-prospatheia-glossologikon-poron-gia-tin-elliniki
6. Plachouras, V. C. (2010). "Archiving the Web sites of Athens University of Economics and Business". *19th Greek Academic Library Conference*. Athens. Ανάκτηση από https://core.ac.uk/download/pdf/38297558.pdf

Ευρετήριο εικόνων

Παράρτημα - Συντελεστές έργου

Δήμητρα Χιώτη, Συντονίστρια	Εθνική Βιβλιοθήκη της Ελλάδος
Σύλβια Πουλημένου	Εθνική Βιβλιοθήκη της Ελλάδος
Ελίζα Μακρίδου	Εθνική Βιβλιοθήκη της Ελλάδος
Δημήτρης Λακιώτης	Εθνική Βιβλιοθήκη της Ελλάδος
Αγγελική Δημητρομανωλάκη	Εθνική Βιβλιοθήκη της Ελλάδος
Μιχάλης Βαζιργιάννης, Επιστημονικός Υπεύθυνος Έργου	Οικονομικό Πανεπιστήμιο Αθηνών
Πολύκαρπος Μελαδιανός	Οικονομικό Πανεπιστήμιο Αθηνών
Σταμάτης Ούτσιος	Οικονομικό Πανεπιστήμιο Αθηνών
Αντώνης Σκανδάλης	Οικονομικό Πανεπιστήμιο Αθηνών
Χρήστος Ξυπολόπουλος	Οικονομικό Πανεπιστήμιο Αθηνών

Γεωργία Αγγελάκη, Διαχειρίστρια Δράσης 2 στο πλαίσιο του «Προγράμματος Υλοποίησης Μετεγκατάστασης της ΕΒΕ στο ΚΠΙΣΝ, 2015-2018»	Φίλοι και Υποστηρικτές ΚΠΙΣΝ
Ουρανία Κουσκουμφεκάκη	Εξωτερική Συνεργάτιδα
Χλόη Παπαδοπούλου	Εξωτερική Συνεργάτιδα
Άννα Μάστορα	Γενική Γραμματεία Επικοινωνίας και Ενημέρωσης
Μαρίνος Παπαδόπουλος	Νομικός Σύμβουλος