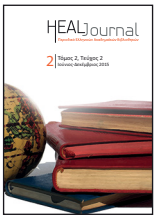
HEALJournal
Περιοδικό Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

Εξαμηνιαία Επιστημονική Έκδοση
Συνδέσμου Ελληνικών Ακαδημαϊκών
Βιβλιοθηκών (ΣΕΑΒ)
http://healjournal.seab.gr



Εξαγωγή Φράσεων από κείμενα: Εφαρμογές και Πλεονεκτήματα για τις ελληνικές Βιβλιοθήκες

Υ. Διδ. Ελένη Θ. Γιαννοπούλου[α]

[α]Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Εθνικό Μετσόβιο Πολυτεχνείο, Ηρώων Πολυτεχνείου 9, Ζωγράφου 15780, Αθήνα, Ελλάδα

Περίληψη

Η «Εξαγωγή Φράσεων» (Keyterm Extraction) από κείμενα, ενώ αποτελεί μια ιδέα που εδώ και αρκετά χρόνια έχει απασχολήσει την ερευνητική κοινότητα, έχει διαμορφωθεί κάτω από διαφορετικό πρίσμα τα τελευταία χρόνια λόγω της έλευσης και της μεγάλης ανάπτυξης του Σημασιολογικού Ιστού. Οι μέθοδοι εξαγωγής φράσεων από κείμενα παρουσιάζουν μεγάλη ποικιλομορφία και μπορούν να ενσωματωθούν σε ένα πλήθος ερευνητικών πεδίων παρέχοντας εφαρμογές που έχουν ως στόχο την παροχή αξιοποιήσιμων υπηρεσιών προς τους τελικούς χρήστες. Στο παρόν άρθρο, αρχικά γίνεται μια εισαγωγή στον χώρο της «Εξαγωγής Φράσεων» από κείμενα/πηγές και εν συνεχεία παρουσιάζονται οι βασικές μέθοδοι που χρησιμοποιούνται επιτυχώς σε ένα πλήθος εφαρμογών προκειμένου να επιτευχθεί ο παραπάνω στόχος. Σε επόμενη ενότητα του παρόντος άρθρου παρουσιάζεται ένα πλήθος διαφορετικών εφαρμογών όπου η «Εξαγωγή Φράσεων» έχει υιοθετηθεί με επιτυχία, εξετάζονται τα πλεονεκτήματα που απορρέουν από την ενσωμάτωση τέτοιων μεθόδων αλλά και οι δυνατότητες εφαρμογής στις Ψηφιακές Βιβλιοθήκες.

Λέξεις-κλειδιά: εξαγωγή φράσεων, εμπλουτισμός μεταδεδομένων, ψηφιακές βιβλιοθήκες, RDF, δεσμευμένα λεξιλόγια, σημασιολογικός ιστός

doi:

Abstract

"Keyterm Extraction", while it is an idea that for many years has interested the research community, it has been formed under a different angle in recent years because of the advent and wide development of the Semantic Web. Methods extracting phrases from texts are very diverse and can be integrated into a multitude of research areas by providing applications that are designed to provide useable services to end users. In this article, which is an introduction in the field of "Keyterm Extraction" texts/sources basic methods are analyzed in order to achieve this objective successfully in a variety of applications. In another section of this article a number of different applications are presented where "Keyterm Exctraction" methods have been adopted successfully. Then the advantages of incorporating such methods and application options in Digital Libraries are also addressed.

Keywords: keyterm exctraction, metadata enrichment, digital libraries, RDF, controlled vocabularies, semantic web

1. Εισαγωγή

Η έλευση του Παγκοσμίου Ιστού αλλά και οι καινοτομίες που η υπηρεσία αυτή εισήγαγε έχουν αλλάξει άρδην τον τρόπο πρόσβασης σε δεδομένα και εφαρμογές. Η υποδομή του Διαδικτύου καλείται πλέον να διαχειριστεί έναν συνεχώς αυξανόμενο όγκο πληροφοριών, οι οποίες δημοσιεύονται με ποικίλους τρόπους και χρησιμοποιούνται για ετερογενείς σκοπούς. Μία απ' αυτές τις καινοτόμες υπηρεσίες αποτελεί και η διάθεση των τεκμηρίων μιας βιβλιοθήκης σε ψηφιακή μορφή μέσω του Παγκοσμίου Ιστού. Η υπηρεσία αυτή έχει καταστήσει άμεσα προσβάσιμα τεκμήρια βιβλιοθηκών ανά τον κόσμο, δίνοντας πρόσβαση σ' αυτά στους χρήστες του Διαδικτύου, χωρίς τον περιορισμό τοποθεσίας που θέτουν οι κλασικές βιβλιοθήκες. Πέρα από τα παραπάνω οφέλη, στις μέρες μας η ανάπτυξη των Ψηφιακών Βιβλιοθηκών (Digital Libraries) έχει συνεισφέρει και στη διάθεση αλλά και διάχυση της ερευνητικής παραγωγής ενός Ιδρύματος με έναν αποδοτικό τρόπο.

Η διαδικασία ψηφιοποίησης και διάθεσης των υπαρχόντων τεκμηρίων αλλά και η διαχείριση των νεοεισερχομένων στην εκάστοτε Ψηφιακή Βιβλιοθήκη αποτελεί μια διαδικασία με πολλά και διαφορετικά βήματα τα οποία δεν στερούνται περιορισμών. Μια βασική εργασία η οποία συντελείται σε οποιαδήποτε Ψηφιακή Βιβλιοθήκη κατά την εισαγωγή μιας νέας εργασίας ή δημοσίευσης σε ένα αποθετήριο αποτελεί η ταυτοποίηση του συγγραφέα του υποκείμενου τεκμηρίου, η οποία έχει ως σκοπό τον συσχετισμό ενός συγγραφέα με μια εργασία/δημοσίευση (Tan, Kan & Lee, 2006). Κατά τη διαδικασία αυτή είναι δυνατόν να ανακύψουν αρκετές δυσκολίες στη διαδικασία εισαγωγής του τεκμηρίου σε ένα σύστημα διαχείρισης βιβλιοθήκης. Οι δυσκολίες αυτές μπορεί να οφείλονται σε ποικίλους παράγοντες, όπως η συνωνυμία, η πολυσημία και η αμφισημία, οι οποίοι που μπορεί να οδηγήσουν ακόμα και σε λανθασμένη εισαγωγή του ονόματος του συγγραφέα στο σύστημα διαχείρισης από τον υπάλληλο της βιβλιοθήκης που είναι υπεύθυνος για τη διαδικασία αυτή.

Εξίσου σημαντική εργασία σε μια Ψηφιακή Βιβλιοθήκη αποτελεί και ο εμπλουτισμός ενός τεκμηρίου με μεταδεδομένα (metadata), διαδικασία που γίνεται κατά κύριο λόγο κατά την καταχώριση ενός τεκμηρίου σε μια Ψηφιακή Βιβλιοθήκη ή σε ένα Ιδρυματικό Αποθετήριο (εφεξής ΙΑ). Η διαδικασία αυτή, λόγω του ότι αποτελείται από πολλά διακριτά βήματα (όπως είναι η καταχώριση μεταδεδομένων κλπ.) μέχρι την επιτυχή καταχώριση του τεκμηρίου στο σύστημα, έχει το μειονέκτημα ότι απαιτεί εξειδικευμένο προσωπικό και είναι ταυτόχρονα χρονοβόρα. Επιπρόσθετα, το κόστος της παραπάνω διαδικασίας ανεβαίνει σε δυσθεώρητα ύψη, όσο ο αριθμός των τεκμηρίων σε μια ψηφιακή βιβλιοθήκη αυξάνεται, καθιστώντας τη διαδικασία αυτή μη – κλιμακώσιμη και μη – χρηστική για τις Ψηφιακές Βιβλιοθήκες και τις συλλογές που περιέχονται σ' αυτές (Han et al., 2003).

Τα τελευταία χρόνια, με τη μεγάλη ανάπτυξη που έχει γνωρίσει ο λεγόμενος «Ιστός των Δεδομένων» (Web of Data) ή «Σημασιολογικός Ιστός» (Semantic Web) έχουν παρουσιαστεί ευκαιρίες ενσωμάτωσης των τεχνολογιών του και εκμετάλλευσης των καινοτομιών του από Ψηφιακές Βιβλιοθήκες ώστε να παρέχουν υπηρεσίες προστιθέμενης αξίας προς τους χρήστες. Οι τεχνολογίες του Σημασιολογικού Ιστού σε συνδυασμό με τις μεθόδους Εξαγωγής Φράσεων και άλλων λέξεων-κλειδιών από κείμενα μπορούν να δώσουν νέα ώθηση στις Ψηφιακές Βιβλιοθήκες, δίνοντας τη δυνατότητα χρήσης των δεδομένων τους με διαφορετικούς αλλά ταυτόχρονα και αποδοτικούς τρόπους, παρέχοντας έτσι στους χρήστες τους καινοτόμες υπηρεσίες.

Στο παρόν άρθρο θα παρουσιαστεί η Εξαγωγή Φράσεων και λέξεων-κλειδιών από κείμενα/πηγές, καθώς και οι μέθοδοι στις οποίες αυτή βασίζεται (Ενότητα 2). Ακόμα, θα αναλυθούν κάποιες από τις πιο αντιπροσωπευτικές εφαρμογές των τεχνικών Εξαγωγής Φράσεων και λέξεων-κλειδιών από κείμενα (Ενότητα 3). Στη συνέχεια θα γίνει μια αναφορά στα πλεονεκτήματα που οι μέθοδοι αυτές μπορούν να προσφέρουν στις Ψηφιακές Βιβλιοθήκες και θα προταθούν ιδέες ενσωμάτωσης των τεχνικών αυτών σε υπάρχουσες πηγές των ελληνικών, ακαδημαϊκών και μη, βιβλιοθηκών (Ενότητα 4) με χρήση των τεχνολογιών του Σημασιολογικού Ιστού.

2. Εξαγωγή φράσεων και λέξεων-κλειδιών από κείμενα

Η τρέχουσα ενότητα συνιστά μια εισαγωγή στο ερευνητικό πεδίο της Εξαγωγής Φράσεων από κείμενα, των μεθόδων που χρησιμοποιούνται προκειμένου να εξαχθεί ένα όσο το δυνατόν πιο σχετικό σύνολο φράσεων, αλλά και των πλεονεκτημάτων ή/και των περιορισμών που οι μέθοδοι αυτές δύνανται να προσφέρουν σε απλούς χρήστες αλλά και στο εξειδικευμένο προσωπικό των βιβλιοθηκών.

Στον τομέα των Ψηφιακών Βιβλιοθηκών η λέξη-κλειδί (keyword) καθορίζεται ως μία ή περισσότερες λέξεις που περιγράφουν ευκρινώς και με ακρίβεια το θέμα ή ένα μέρος του θέματος που αναφέρεται σε ένα έγγραφο (Feather & Sturges, 1996). Αντίθετα, σύμφωνα με το Oxford Dictionary (2015), στο πεδίο της Επιστήμης των Πληροφοριών ο όρος Keyword περιγράφει μια λέξη που χρησιμοποιείται σε ένα σύστημα ανάκτησης πληροφοριών για να υποδείξει το περιεχόμενο ενός εγγράφου. Ακόμα, σε ένα σύστημα διαχείρισης βάσεων δεδομένων

ο όρος αυτός περιγράφει λέξεις που βρίσκονται σε ένα ευρετήριο. Οι λέξεις αυτές έχουν εισαχθεί από τον άνθρωπο που κάνει την ευρετηρίαση έτσι ώστε να διευκολύνεται η αναζήτηση μέσα σ' αυτό. Έναν ακόμα σχετικό όρο αποτελούν οι φράσεις-κλειδιά (keyphrases), οι οποίες μπορούν να χρησιμοποιηθούν για να χαρακτηρίσουν το περιεχόμενο ενός εγγράφου χρησιμοποιώντας φράσεις που μπορεί να αποτελούνται από περισσότερες από μία λέξεις. Βασική διαφορά των φράσεων από τις λέξεις-κλειδιά αποτελεί το γεγονός πως μια λέξη-κλειδί πρέπει υποχρεωτικά να εμφανίζεται μέσα στο σώμα του κειμένου, ενώ οι φράσεις δεν είναι απαραίτητο να εμφανίζονται εκεί. Για τον λόγο αυτό και έχει αποδειχθεί σε ένα πλήθος περιπτώσεων πως η χρήση των φράσεων έναντι των λέξεων-κλειδιών μπορεί να φέρει πιο σχετικά αποτελέσματα σε μια αναζήτηση (Mendelyan, 2005). Για τον ίδιο λόγο, οι βιβλιοθηκονόμοι χρησιμοποιούν θεματικές επικεφαλίδες (subject headings). Οι θεματικές επικεφαλίδες αποτελούν όρους ή φράσεις που περιγράφουν το περιεχόμενο ενός τεκμηρίου και, επιπρόσθετα, μπορούν να χρησιμοποιηθούν για να ομαδοποιήσουν και να οργανώσουν τις συλλογές μιας βιβλιοθήκης (Mendelyan, 2005). Για παράδειγμα, έγγραφα τα οποία σχετίζονται με το ίδιο θέμα μπορούν να ομαδοποιηθούν έτσι ώστε να υπάρχει σ' αυτά γρήγορη πρόσβαση με την επιλογή συγκεκριμένων θεματικών επικεφαλίδων.

Η έννοια της Εξαγωγής Φράσεων (keyphrase extraction) από κείμενα αναφέρεται στην εξαγωγή από κάποιο κείμενο μιας συγκεκριμένου μεγέθους λίστας φράσεων (keyphrase list), στην οποία καθεμία φράση μπορεί να αποτελείται από δύο ή περισσότερες λέξεις αντί για μια μεμονωμένη λέξη κλειδί (Turney, 1999). Στόχος αυτής της λίστας είναι να αποτυπώσει τα βασικά σημεία ή τις βασικές έννοιες ενός δεδομένου εγγράφου. Σε σχέση με την ευρετηρίαση πλήρους κειμένου (full-text indexing) όπου όλοι οι όροι που εμφανίζονται στο κείμενο μπορεί στη συνέχεια να αποθηκεύονται σε κάποια μονάδα αποθήκευσης, όπως σε μια βάση δεδομένων, η ευρετηρίαση με βάση λέξεις ή φράσεις-κλειδιά (keyterm indexing) απαιτεί λιγότερο αποθηκευτικό χώρο και παρέχει μια συνοπτική επισκόπηση των θεματικών περιοχών από μια συλλογή ή ένα μεμονωμένο κείμενο.

Όπως οι θεματικές επικεφαλίδες έτσι και οι λέξεις-κλειδιά, όσο και σχετικές με το θέμα φράσεις, μπορούν εναλλακτικά να επιλεγούν από ένα τυποποιημένο σύνολο περιγραφέων όπως ένα ελεγχόμενο λεξιλόγιο (controlled vocabulary). Τα ελεγχόμενα λεξιλόγια αποτελούν μια προσεκτικά επιλεγμένη αλληλουχία από λέξεις και φράσεις οι οποίες χρησιμοποιούνται τόσο στον τομέα των Ψηφιακών Βιβλιοθηκών όσο και γενικότερα στο πεδίο της Επιστήμης των Πληροφοριών για την επισημείωση τμημάτων πληροφορίας που μπορεί να αποτελούν κομμάτι μιας ιστοσελίδας ή ενός εγγράφου/τεκμηρίου, έτσι ώστε τα παραπάνω να μπορούν να ανακτηθούν με έναν εύκολο τρόπο στα πλαίσια μιας αναζήτησης. Ένας ακόμα τρόπος για την ανάθεση λέξεων ή φράσεων κλειδιών αποτελεί η εξαγωγή τους από το σώμα του κειμένου, από τον τίτλο ή και από την περίληψη αυτού (Kim, S. N., 2009). Σ' αυτή την εργασία, αξιολογήθηκε ο αντίκτυπος που έχουν διαφορετικές μέθοδοι στην κατηγοριοποίηση κειμένου με βάση την εξαγωγή τους από ένα δεδομένο τμήμα κείμενου. Επιλέχθηκαν, δηλαδή, οι λέξεις του τίτλου, οι πρώτες λέξεις του κειμένου αλλά και η περίληψή του. Σ' αυτή την περίπτωση καταδεικνύεται πως οι μέθοδοι εξαγωγής φράσεων με χρήση συγκεκριμένων τμημάτων του δεδομένου κειμένου μπορεί να λειτουργούν καλύτερα στην πλειοψηφία των εξεταζόμενων περιπτώσεων σε σχέση με τη χρήση ολόκληρου του κειμένου.

Μια βασική χρήση της λίστας των φράσεων αποτελεί η δημοσίευση ενός κειμένου σε ένα επιστημονικό περιοδικό, όπου ζητείται από τους συγγραφείς, εκτός από το κείμενο της δημοσίευσης και μια λίστα με αντιπροσωπευτικές φράσεις ή λέξεις-κλειδιά. Συνήθως, οι φράσεις αυτές είτε παρέχονται απευθείας από τους συγγραφείς και στη συνέχεια επιθεωρούνται από το εξειδικευμένο προσωπικό των βιβλιοθηκών είτε η διαδικασία αυτή γίνεται εξ ολοκλήρου από το προσωπικό, το οποίο είναι υπεύθυνο για την κατάρτιση της λίστας αυτής. Η διαδικασία αυτή συνήθως είναι επίπονη και χρονοβόρα, ενώ, επίσης, μειονεκτεί στο γεγονός ότι είναι χειρωνακτική, και άρα υπόκειται σε λάθη. Συμπληρωματικά, θα μπορούσαμε να πούμε ότι είναι μια υποκειμενική διαδικασία στην οποία πολύ σημαντικό ρόλο παίζει ο ανθρώπινος παράγοντας, διότι είναι λογικό κάθε άνθρωπος να καταρτίσει διαφορετική λίστα φράσεων. Η παραπάνω περίπτωση αποτελεί μία από τις πολλές περιπτώσεις στις οποίες η εφαρμογή μεθόδων Εξαγωγής Φράσεων μπορεί να δώσει λύση.

Λόγω των προβλημάτων/περιορισμών των τεχνικών της Εξαγωγής Φράσεων, η τελευταία έχει συγκεντρώσει αρκετά μεγάλο ερευνητικό ενδιαφέρον. Υπάρχουν πολλές προσεγγίσεις που εφαρμόζουν διαφορετικές μεθόδους χρησιμοποιώντας διαφορετικούς τύπους κειμένων ή ακόμα και συγκεκριμένα τμήματα των δεδομένων κειμένων έτσι ώστε να καταρτίσουν μια λίστα αντιπροσωπευτικών φράσεων με έναν αυτόματο ή ημιαυτόματο τρόπο, προκειμένου να υποβοηθήσουν το έργο συγγραφέων ή/και βιβλιοθηκονόμων.

Στο σημείο αυτό μπορούμε να πούμε ότι το παραπάνω πρόβλημα έγκειται στην αναγωγή της εξαγωγής φράσεων από κείμενα με *χειροκίνητο* τρόπο σε εξαγωγή φράσεων με *αυτόματο* ή *ημιαυτόματο* τρόπο. Ένα ακόμα πλεονέκτημα της χρήσης των φράσεων κλειδιών είναι η πολύ-λειτουργικότητα (multi-functionality) που τις διακρίνει. Οι φράσεις-κλειδιά αλλά και τα σύνολα φράσεων μπορούν να χρησιμοποιηθούν σε ποικίλες

εφαρμογές επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) όπως κατηγοριοποίηση κειμένου (text clustering) και ταξινόμηση (classification) (Jones & Mahoui, 2000).

Δύο ακόμα πεδία όπου οι φράσεις-κλειδιά μπορούν να εφαρμοστούν με επιτυχία αποτελούν η ανάκτηση πληροφοριών που βασίζεται στο περιεχόμενο (content-based retrieval) και η θεματική αναζήτηση (topic search). Τέλος, όπως αναφέρθηκε και παραπάνω, οι φράσεις μπορούν να χρησιμοποιηθούν αποδοτικά κατά την αναζήτηση (Hulth, 2004) ή/και την πλοήγηση είτε στον Παγκόσμιο Ιστό είτε πιο συγκεκριμένα στα στενά όρια μιας Ψηφιακής Βιβλιοθήκης. Ένας παράγοντας που διαδραματίζει πολύ σημαντικό ρόλο στη συγκεκριμένη περίπτωση είναι η ποιότητα των φράσεων που θα εξαχθούν από το κείμενο. Για τον λόγο αυτό, αποτελεί σημαντικό στοιχείο η εξαγωγή να έχει οδηγήσει στην παραγωγή ποιοτικών και ταυτόχρονα συγκεκριμένων περιγραφέων, οι οποίοι, σε περίπτωση που η εξαγωγή αυτών δεν γίνεται με αυτόματο ή ημιαυτόματο τρόπο, θα παρέχονται από ανθρώπους ειδικούς στο αντικείμενο.

Οι φράσεις μπορούν ακόμα να χρησιμοποιηθούν για τη δημιουργία ενός ευρετηρίου. Σ' αυτή την περίπτωση, ένας χρήστης είναι δυνατόν να ανακαλύψει μέσω της αναζήτησης είτε στον Ιστό είτε σε μια Ψηφιακή Βιβλιοθήκη ένα άρθρο που μπορεί να καλύπτει τις ανάγκες του μέσα σε μικρό χρονικό διάστημα. Επίσης, είναι δυνατή η χρήση της λίστας αυτής κατά τη διάρκεια μιας αναζήτησης. Οι εν λόγω φράσεις μπορούν να αποτελέσουν την είσοδο μιας μηχανής αναζήτησης αντί για μεμονωμένες λέξεις-κλειδιά, έτσι ώστε η τελευταία να επιστρέψει πιο σχετικά ή/και πιο ακριβή αποτελέσματα. Τα αποτελέσματα της αναζήτησης μπορούν να βελτιστοποιηθούν περαιτέρω εάν αυτή η λίστα των φράσεων συνδυαστεί ή κάποιες φράσεις που ανήκουν στη λίστα αντιστοιχιστούν με όρους που προέρχονται από κάποιο ελεγχόμενο λεξιλόγιο, όπως αναφέρθηκε σε προηγούμενη παράγραφο. Έτσι, μπορεί να καταστεί δυνατή η ανάκτηση περισσότερο σχετικών ή πιο γενικών ή ακόμα και πιο ειδικών αποτελεσμάτων, λόγω του ότι τα ελεγχόμενα λεξιλόγια ή οι θησαυροί ενσωματώνουν συνήθως ένα είδος ιεραρχίας. Στα υψηλότερα επίπεδα της ιεραρχίας υπάρχουν οι πιο γενικοί όροι, ενώ στα ενδιάμεσα και στα χαμηλότερα οι πιο ειδικοί. Έτσι είναι δυνατόν να τροποποιηθεί η αναζήτηση επιλέγοντας από την ιεραρχία είτε πιο εδικούς είτε πιο γενικούς όρους ανάλογα πάντα με τα αποτελέσματα που επιστρέφονται. Απόρροια αυτού είναι στις περισσότερες περιπτώσεις μια μικρότερη σε μέγεθος και ταυτόχρονα πιο σχετική λίστα αποτελεσμάτων, δηλαδή με περισσότερο ποιοτικά αποτελέσματα. Όπως είναι φανερό από τα παραπάνω, παρόλο που οι φράσεις αυτές μπορεί να χρησιμοποιούνται για να επιτελέσουν διαφορετικούς σκοπούς και να χρησιμοποιηθούν από διαφορετικές εφαρμογές, υπάρχει μια κοινή απαίτηση: μια μικρή λίστα από φράσεις οι οποίες θα αποτυπώνουν το βασικό νόημα ενός τεκμηρίου ή εγγράφου (Turney, 1999).

Οι υπάρχουσες προσεγγίσεις ευρετηρίασης με βάση φράσεις-κλειδιά μπορούν να χωριστούν σε δύο μεγάλες κατηγορίες (Mendelyan, 2005). Μπορούν να θεωρηθούν είτε ως μια διαδικασία εξαγωγής (extraction) είτε ως μια διαδικασία ανάθεσης (assignment). Στην πρώτη περίπτωση, οι όροι και οι φράσεις που εμφανίζονται στο σώμα του κειμένου αναλύονται με σκοπό τελικά να επιλεγούν οι πιο αντιπροσωπευτικές. Στη δεύτερη περίπτωση, τα έγγραφα κατηγοριοποιούνται σε έναν προκαθορισμένο αριθμό ομάδων με βάση αυτές τις φράσεις. Η εργασία της Mendelyan (2005), επικεντρώνεται στον καθορισμό των πιο αντιπροσωπευτικών φράσεων και όχι λέξεων-κλειδιών ως περιγραφέων για το περιεχόμενο ενός κειμένου. Οι φράσεις αυτές μπορεί να έχουν προκύψει με τη χρήση ενός ελεγχόμενου λεξιλογίου και επιλέγονται, τελικά, με τη βοήθεια ενός από τους πιο γνωστούς αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται για την εξαγωγή φράσεων από κείμενα, του KEA (Witten Paynter, Frank, Gutwin, Nvill-Manning, 1999). Οι φράσεις οι οποίες επιλέγονται στο τελικό στάδιο μπορεί να αποτελούν φράσεις που εμφανίζονται μέσα στο κείμενο αλλά όχι απαραιτήτως, και εκεί έγκειται και η διαφοροποίηση σε σχέση με τη χρήση λέξεων ως περιγραφέων, καθώς οι τελευταίες πρέπει να εμφανίζονται μέσα στο υπό εξέταση κείμενο. Σύμφωνα με την παραπάνω εργασία, οι φράσεις αυτές αποτελούν χρήσιμα μεταδεδομένα τόσο για φυσικές όσο και για ψηφιακές βιβλιοθήκες, καθώς υποβοηθούν την οργάνωση των συλλογών που υπάρχουν σε μια βιβλιοθήκη με βάση το περιεχόμενο των εγγράφων. Στη συγκεκριμένη εργασία (Mendelyan, 2005) προτείνεται μια νέα έκδοση του αλγορίθμου KEA, ο αλγόριθμος KEA++, οποίος συνδυάζει τις δύο περιπτώσεις της εξαγωγής φράσεων και της ανάθεσης σε μια συνολική διαδικασία, όπου οι όροι και οι φράσεις που εξάγονται από τα κείμενα θα πρέπει να είναι μέρος κάποιου ελεγχόμενου λεξιλογίου ή οντολογίας. Τέλος, προτείνουν η διαδικασία ευρετηρίασης να επεκταθεί με την ανάλυση των σημασιολογικών πληροφοριών που σχετίζονται με τους φράσεις που εξάγονται από τα αρχεία, για παράδειγμα μέσω της συσχέτισης των φράσεων που υπάρχουν στο αρχείο με τους όρους που υπάρχουν στο επιλεγμένο λεξιλόγιο.

Οι δύο παραπάνω προσεγγίσεις (KEA, KEA++) βασίζουν ένα τμήμα της υλοποίησής τους στους αλγορίθμους αποκατάληξης (stemming algorithms) Porter (1980) και Lovins (1968), οι οποίοι είναι οι δύο πιο δημοφιλείς αλγόριθμοι για την αφαίρεση καταλήξεων από αγγλικές λέξεις. Και οι δύο αυτοί αλγόριθμοι χρησιμο-

ποιούν ευρετικούς κανόνες για την αφαίρεση ή τον μετασχηματισμό των καταλήξεων ενός δεδομένου κειμένου ή τμήματος αυτού. Σ' αυτό το σημείο θα πρέπει να σημειωθεί πως υπάρχει και η εναλλακτική προσέγγιση της χρήσης ενός λεξιλογίου που θα απαριθμεί ρητά το λήμμα για κάθε λέξη που θα μπορούσε να προκύψει σε ένα δεδομένο κείμενο. Παρόλα αυτά, προτιμάται η χρήση των ευρετικών κανόνων έναντι του λεξιλογίου, λόγω της κοπιαστικής προσπάθειας που απαιτείται για την κατασκευή του λεξιλογίου και της υπολογιστικής ικανότητας που απαιτείται για να είναι αποδοτική η χρήση του. Όσο αφορά τη λειτουργία τους, έχει τεκμηριωθεί πως ο αλγόριθμος του Lovins είναι περισσότερο επιθετικός απ' αυτόν του Porter. Επίσης, ο Lovins είναι περισσότερο πιθανό να αντιστοιχίσει δύο λέξεις που έχουν την ίδια ρίζα, αλλά έχει και μεγαλύτερη ανοχή σε λάθη (Turney, 1999). Ακόμα, έχει παρατηρηθεί πως η επιθετική αφαίρεση, όπως αυτή που συμβαίνει στον αλγόριθμο του Lovins, είναι καλύτερη για την εξαγωγή φράσεων από κείμενα, σε σχέση με τη συντηρητική αφαίρεση.

Η διαδικασία που περιγράφηκε παραπάνω μπορεί να έχει πολλά οφέλη και εφαρμογές στις Ψηφιακές Βιβλιοθήκες με ευεργετικά αποτελέσματα για τους τελικούς χρήστες, καθώς θα μπορούσε να χρησιμοποιηθεί σε ένα σύστημα προτάσεων λέξεων ή φράσεων-κλειδιών κάθε φορά που ένας συγγραφέας καταθέσει μια νέα δημοσίευση ή κάποιο άλλο τεκμήριο. Η διαδικασία αυτή θα μπορούσε ακόμα να χρησιμοποιηθεί ως ένας ημι-αυτόματος τρόπος για να κατευθύνει την επιλογή φράσεων είτε από τους συγγραφείς είτε από το προσωπικό των βιβλιοθηκών.

Για όλους τους παραπάνω λόγους, κρίνεται αναγκαία η ανάπτυξη εργαλείων αυτόματης παραγωγής φράσεων από κείμενα. Παρόλη τη φανερή ανάγκη που υπάρχει σε διάφορους τομείς έρευνας, οι φράσεις αυτές τείνουν να χρησιμοποιούνται για την επισημείωση ενός πολύ μικρού αριθμού εγγράφων που είναι διαθέσιμα στον Ιστό.

3. Εφαρμογές

Με την επέκταση και ευρύτατη χρήση του Διαδικτύου και των εταιρικών ενδοδικτύων, η διαχείριση αρχείων διαφορετικών μορφότυπων είναι ύψιστης σημασίας. Πολλοί ερευνητές θεωρούν πως η παροχή ποιοτικών μεταδεδομένων είναι βασικός παράγοντας της επιτυχίας στη διαχείριση, ανάκτηση και αναζήτηση των αρχείων. Τα μεταδεδομένα αποτελούν μετά-πληροφορίες που αφορούν ένα έγγραφο ή ένα σύνολο από έγγραφα. Ειδικότερα στον βιβλιογραφικό τομέα, υπάρχουν διάφορα πρότυπα που αφορούν τα μεταδεδομένα αυτά, όπως το Dublin Core Metadata Element Set, ή διάταξη δεδομένων σύμφωνα με το πρότυπο MARC (Machine-Readable Cataloguing). Μία από τις πλέον βασικές κατηγορίες εφαρμογών των φράσεων-κλειδιών αφορά τη χρήση αυτών ως μεταδεδομένων ενός εγγράφου, και μπορούν επιτυχώς να χρησιμοποιηθούν στη διαχείριση αυτού. Τα πρότυπα αυτά μεταδεδομένων έχουν προδιαγράψει ένα σύνολο πεδίων για την αποθήκευση φράσεων που αφορούν το υποκείμενο τεκμήριο. Κάθε τεκμήριο που κατατίθεται σε μια Ψηφιακή Βιβλιοθήκη είναι απαραίτητο να συνοδεύεται και από ένα σύνολο μεταδεδομένων που να το περιγράφουν με έναν αποδοτικό τρόπο, επιτρέποντας έτσι την ευρετηρίαση και τελικά την ανάκτησή του από τους τελικούς χρήστες. Επειδή αυτή η διαδικασία βαραίνει συνήθως είτε τον συγγραφέα είτε τον βιβλιοθηκονόμο, θεωρούμε πως θα ήταν πολύ βοηθητικό εάν υπήρχε μια αυτόματη διαδικασία παραγωγής μεταδεδομένων που να αφορούν το τεκμήριο. Οι φράσεις αυτές μπορεί να έχουν προκύψει από κάποιο ελεγχόμενο λεξιλόγιο ή θησαυρό και να χρησιμοποιηθούν για τον εμπλουτισμό των μεταδεδομένων ενός τεκμηρίου υπερβαίνοντας, προβλήματα υποκειμενικότητας που υπεισέρχονται στη διαδικασία λόγω της εμπλοκής του ανθρώπινου παράγοντα. Τέλος, η διαδικασία αυτή θα μπορούσε να χρησιμοποιηθεί ως ένας ημι-αυτόματος τρόπος για να κατευθύνει την επιλογή φράσεων είτε από τους συγγραφείς είτε από το προσωπικό των βιβλιοθηκών.

Η χρήση των φράσεων αυτών, όπως διαφαίνεται και από τα προαναφερθέντα, μπορεί να επιτελέσει πολλούς και διαφορετικούς σκοπούς. Έναν από αυτούς αποτελεί και η σύνοψη του θέματος (summarization) ενός επιστημονικού περιοδικού, ενός τεκμηρίου ή πιο απλά ενός κειμένου ή τμήματος αυτού. Έτσι, είναι δυνατόν να δημιουργηθεί, είτε από τον ίδιο τον συγγραφέα είτε από το προσωπικό, η σύνοψη του υποκείμενου θέματος εφαρμόζοντας τεχνικές Εξαγωγής Φράσεων, το αποτέλεσμα των οποίων θα τοποθετηθεί στην αρχή του κειμένου, προκειμένου να προϊδεάσει τον αναγνώστη για το περιεχόμενό του. Εν προκειμένω, αυτή η λίστα αυτή των φράσεων μπορεί να λειτουργήσει ως μια «ειδική» μορφή περίληψης (Barker & Cornacchia, 2000). Το πλεονέκτημα που προκύπτει από την παραπάνω εφαρμογή είναι αδιαμφισβήτητο, διότι έτσι δύναται ο εκάστοτε χρήστης να αναγνωρίσει σε ελάχιστο χρόνο εάν το θέμα ενός δεδομένου κειμένου είναι μέσα στα ενδιαφέροντά του ή όχι. Μ' αυτόν τον τρόπο μπορεί να εξοικονομήσει χρόνο κατά την αναζήτηση ενός άρθρου, σε σύγκριση πάντα με τον χρόνο που θα απαιτούσε η αντίστοιχη αναζήτηση εάν πραγματοποιούνταν με χρήση λέξεων και όχι φράσεων.

Μια ακόμα εργασία η οποία έχει κεντρίσει το ενδιαφέρον της επιστημονικής κοινότητας είναι η λεγόμενη επισημείωση του περιεχομένου μιας σελίδας στο Διαδίκτυο ή ενός αρχείου (highlighting). Η διαδικασία αυτή αφορά τον εντοπισμό των βασικών στοιχείων ενός αρχείου ή σελίδας, δηλαδή φράσεων που θα καθορίζουν το περιεχόμενό του. Επίσης, είναι περισσότερο ευδιάκριτα τα βασικά στοιχεία ενός κειμένου

εάν αυτά επισημειωθούν. Αυτή η διαδικασία επιτυγχάνεται δίνοντας έμφαση ή επισημειώνοντας τμήματα κειμένου χρησιμοποιώντας κάποια ειδική γραμματοσειρά ή αλλάζοντας το χρώμα υποβάθρου του κειμένου. Λόγω της ερευνητικής δραστηριότητας που υπάρχει στο πεδίο αυτό, υπάρχουν κάποια προγραμματιστικά εργαλεία που υποβοηθούν τη διαδικασία αυτή επισημειώνοντας αυτόματα τα εν λόγω σημεία του κειμένου. Στη συνέχεια, χρησιμοποιώντας τα τμήματα αυτά είναι δυνατόν να παραχθεί μια περίληψη του κειμένου αυτού που να βασίζεται στις φράσεις που έχουν επισημειωθεί με τη διαδικασία που αναφέρθηκε προηγουμένως. Λόγω του ότι η διαδικασία αυτή εμπεριέχει μεγάλο βαθμό εξάρτησης από τον ανθρώπινο παράγοντα, θα ήταν πολύ αποδοτικό εάν υπήρχε ένας αυτόματος τρόπος παραγωγής των ως άνω φράσεων που θα μπορούσε να υποβοηθήσει την επιλογή των φράσεων από τους ειδικούς του εκάστοτε πεδίου.

Μία ακόμα βασική εφαρμογή των φράσεων αυτών αποτελεί η χρήση τους για την ευρετηρίαση. Μία αλφαβητική λίστα από φράσεις η οποία έχει προκύψει από μια συλλογή αρχείων ή από τμήματα ενός μακροσκελούς εγγράφου μπορεί να υποστηρίξει το ρόλο ενός ευρετηρίου. Η λίστα αυτή μπορεί να έχει προκύψει με χρήση ενός αλγορίθμου όπως ο Gen-Ex (Turney, 1999). Σ' αυτή την περίπτωση μπορεί ένας χρήστης να εισαγάγει μια λέξη ή ένα τμήμα μιας λέξης, ενώ, σε επόμενο βήμα, το σύστημα μπορεί να παράξει αυτόματα, με τη μορφή λίστας, φράσεις οι οποίες βασίζονται στην λέξη που εισήγαγε ο χρήστης. Αυτές μπορεί απλώς να έχουν την ίδια ρίζα ή ακόμα και να είναι συνώνυμες με την αρχική. Επιλέγοντας την κατάλληλη φράση, το σύστημα επιστρέφει στον χρήστη τη λίστα με όλα τα άρθρα που είναι σχετικά μ' αυτή τη φράση.

Ακόμα, είναι δυνατή η χρήση αυτών ως όρων αναζήτησης σε μια μηχανή αναζήτησης για τη βελτιστοποίηση ενός ερωτήματος (query refinement). Η χρήση μιας μηχανής αναζήτησης, είναι συχνά μια επαναληπτική διαδικασία. Κατά τη διαδικασία μιας «κλασικής» αναζήτησης ο εκάστοτε χρήστης χρησιμοποιεί λέξεις-κλειδιά για να πραγματοποιήσει μια αναζήτηση γύρω από το θέμα που τον ενδιαφέρει, ανακτώντας νέες πληροφορίες σε κάθε νέο έγγραφο/αρχείο που ανακαλύπτει, ενώ ταυτόχρονα, συνδυάζει τις πληροφορίες αυτές έτσι ώστε να βελτιστοποιήσει την αναζήτησή του χρησιμοποιώντας νέες λέξεις ή φράσεις. Υπ' αυτή την έννοια, η διαδικασία αυτή μπορεί να απλοποιηθεί χρησιμοποιώντας νέες προγραμματιστικές διεπαφές που ενσωματώνουν τεχνολογίες Εξαγωγής Φράσεων από κείμενα και έχουν δημιουργηθεί με σκοπό να αυτοματοποιήσουν, κατά το δυνατό, την παραπάνω διαδικασία. Υπάρχουν διεπαφές που έχουν υλοποιηθεί και υποστηρίζουν την επαναληπτική βελτιστοποίηση ερωτημάτων. Η λειτουργία μιας τέτοιας προγραμματιστικής διεπαφής παρουσιάζεται στο άρθρο του Turney (1999), όπου ένας χρήστης θέτει ένα ερώτημα στη μηχανή αναζήτησης και του επιστρέφονται άρθρα σχετικά με το ερώτημα που έθεσε, αλλά ταυτόχρονα του παρέχει και προτάσεις για να περιορίσει την αναζήτησή του. Οι όροι αναζήτησης που θέτει ο χρήστης συνδυάζονται μεταξύ τους, έτσι ώστε η λίστα των αποτελεσμάτων να μικραίνει με κάθε επανάληψη της αναζήτησης.

Μία ακόμα εφαρμογή των μεθόδων αυτόματης Εξαγωγής Φράσεων αποτελεί η ανάλυση των προτύπων χρήσης (usage patterns) στα αρχεία ενός διακομιστή Ιστού (web server logs). Αποτελεί μια ιδιαίτερη εφαρμογή που βασίζεται στην ανάγκη που έχουν οι διαχειριστές ενός διακομιστή Ιστού να γνωρίζουν τι συνήθως αναζητούν οι επισκέπτες των ιστοσελίδων τους. Οι πληροφορίες αυτές ως επί το πλείστον αποθηκεύονται στα λεγόμενα αρχεία καταγραφής Ιστού (log files). Έχουν υλοποιηθεί πλέον και διατίθενται εμπορικά αρκετά προϊόντα που κάνουν αυτή την ανάλυση. Συνήθως τα λογισμικά αυτά παράγουν μια περίληψη των προτύπων κίνησης (traffic patterns) και παράγουν και μια λίστα με τα πιο δημοφιλή αρχεία που συνήθως επισκέπτονται οι χρήστες σε έναν διακομιστή Ιστού. Αν συνδυαστεί ένα τέτοιο πρόγραμμα με ένα πρόγραμμα εξαγωγής φράσεων, θα ήταν δυνατόν να χρησιμοποιηθούν οι φράσεις αυτές, έτσι ώστε να παραχθεί μια λίστα με τις πιο συχνά χρησιμοποιούμενες φράσεις. Η λίστα αυτή μπορεί να δώσει στους διαχειριστές της ιστοσελίδας μια ιδέα για το ποια είναι τα πιο δημοφιλή θέματα που αναζητούν οι χρήστες που την επισκέπτονται.

Σε προηγούμενη παράγραφο του παρόντος εγγράφου παρουσιάστηκε ο αλγόριθμος KEA++, ο οποίος πραγματοποιεί ευρετηρίαση εγγράφων με τη βοήθεια ενός θησαυρού (Mendelyan & Witten, 2006). Σ' αυτή την προσέγγιση χρησιμοποιούνται τεχνικές μηχανικής μάθησης και σημασιολογικές πληροφορίες σχετικά με τους όρους που κωδικοποιούνται σε ένα δομημένο ελεγχόμενο λεξιλόγιο. Το βασικότερο πλεονέκτημα της προσέγγισης αυτής έναντι της κλασικής εξαγωγής φράσεων είναι η χρήση ενός ελεγχόμενου λεξιλογίου, το οποίο εξαλείφει την εμφάνιση φράσεων που δεν έχουν νόημα ή είναι λάθος και, επίσης, παρέχει μια αξιοσημείωτη βελτίωση στην επίδοση του αλγορίθμου KEA++. Ένα ακόμα πλεονέκτημα που παρέχει η διαδικασία αυτή έναντι της κλασικής διαδικασίας ανάθεσης όρων, η οποία χρησιμοποιεί ήδη ένα ελεγχόμενο λεξιλόγιο, είναι η απαίτηση για ένα πολύ μικρό πλήθος δεδομένων εκπαίδευσης του υποκείμενου συστήματος. Τέλος, στην εργασία αυτή η διαδικασία αξιολόγησης που εφαρμόστηκε καταδεικνύει ότι η απόδοση του συστήματος είναι ανεξάρτητη από το μέγεθος του ελεγχόμενου λεξιλογίου που χρησιμοποιείται.

4. Ευκαιρίες αξιοποίησης μεθόδων Εξαγωγής φράσεων σε συνδυασμό με τις τεχνολογίες Σημασιολογικού Ιστού στις ελληνικές Βιβλιοθήκες

Αν λάβει κάνεις υπόψη του την πληθώρα των εφαρμογών Εξαγωγής Φράσεων που διατίθενται σε διαφορετικά πεδία εφαρμογής, αλλά και τις εφαρμογές που θα μπορούσαν να προκύψουν χρησιμοποιώντας τις υπάρχουσες διεπαφές, ή ακόμα και συνδυάζοντάς τες με κάποιον τρόπο, είναι φανερό πως υπάρχουν αρκετές ευκαιρίες αξιοποίησης των μεθόδων Εξαγωγής Φράσεων στις ελληνικές Βιβλιοθήκες, ακαδημαϊκές ή μη. Όπως αναλύθηκε σε προγενέστερη παράγραφο του παρόντος άρθρου, λόγω των περιορισμών που θέτει μια κλασική αναζήτηση με λέξεις-κλειδιά υπάρχει η ανάγκη από την πλευρά των χρηστών για περισσότερο εύκολες αναζητήσεις που θα επιστρέφουν πιο σχετικά και ποιοτικά αποτελέσματα.

Μια προσέγγιση προς αυτή την κατεύθυνση είναι η δημιουργία νέων μηχανών αναζήτησης που θα έχουν τη δυνατότητα να πραγματοποιούν κάποιο είδος συλλογισμού (reasoning). Η λύση αυτή, παρόλο που ακολουθείται σε ορισμένες περιπτώσεις (Zimmermann, Lopes, Polleres, Straccia, 2012) με επιτυχία, είναι δύσκολα υλοποιήσιμη και δεν επιφέρει πάντοτε τα επιθυμητά αποτελέσματα. Μια διαφορετική ιδέα στηρίζεται στη λογική ότι δεδομένου ότι υπάρχουν αρκετές δυσκολίες στο να δημιουργηθούν περισσότερο «έξυπνες» μηχανές αναζήτησης ή λύση θα ήταν να εμπλουτιστούν τα μεταδεδομένα των υποκείμενων πηγών αναζήτησης έτσι ώστε να είναι πιο εύκολη η διαδικασία της αναζήτησης. Αυτό περιλαμβάνει τη μετατροπή των αδόμητων πηγών δεδομένων σε νέες δομημένες πηγές, έτσι ώστε η αναζήτηση να γίνεται με πιο αποδοτικό και ευέλικτο τρόπο (Zervanou, Konteontzelous, Bosch, Anania dou, 2011).

Σ' αυτή την ιδέα στηρίζεται και το όραμα του Σημασιολογικού Ιστού (Semantic Web), όπως εκφράστηκε από τον εμπνευστή του Tim Berners-Lee (Berners-Lee, 2001), με τη μετάβαση από έναν Ιστό εγγράφων όπου το έγγραφο συνιστούσε τη βασική δομική μονάδα σε έναν Ιστό αντικειμένων (Web of Data), όπου η σημασία πλέον δίνεται στις έννοιες που «κρύβονται» πίσω από κάθε πηγή δεδομένων. Για να γίνει πραγματικότητα το όραμα αυτό του Σημασιολογικού Ιστού έχουν προταθεί και εφαρμοστεί συγκεκριμένα πρότυπα, όπως το RDF μοντέλο (Schreiber & Raimond, 2014) και η OWL (Hitzler, Krotzsch, Parsia, Patel-Schneider, Rudolph, 2014), γλώσσες επισημείωσης πηγών που παρέχουν όμως διαφορετικά επίπεδα εκφραστικότητας. Το μοντέλο RDF, λόγω της απλότητας που προσφέρει στην περιγραφή και στην αναπαράσταση δεδομένων, αποτελεί βασικό μοντέλο αναπαράστασης του Σημασιολογικού Ιστού. Ένα ακόμα πρότυπο που χρησιμοποιείται για την αναζήτηση είναι η γλώσσα SPARQL (http://www.w3.org/TR/sparql11-overview/), η οποία σε συνδυασμό με τις οντολογίες (ontologies), ολοκληρώνει τη στοίβα τεχνολογιών του Σημασιολογικού Ιστού. Οι οντολογίες ή αλλιώς σύνολα στοιχείων δεδομένων (metadata element sets) (Baker et al., 2011) αποτελούν την περιγραφή μιας Θεματικής περιοχής ή ενός πεδίου εφαρμογής με χρήση ενός φορμαλισμού. Οι οντολογίες χρησιμοποιούνται για την περιγραφή αντικειμένων με τη βοήθεια URIs (Uniform Resource Identifiers). Εκτός από τα σύνολα στοιχείων δεδομένων έχουν αναπτυχθεί και τα λεγόμενα λεξιλόγια τιμών (value vocabularies), τα οποία μπορούν να χρησιμοποιηθούν για την περιγραφή/επισημείωση οποιασδήποτε πηγής δεδομένων όπως άτομα, γεωγραφικές τοποθεσίες, προϊόντα κλπ. Στην κατηγορία των λεξιλογίων τιμών εμπίπτουν, επίσης, οι θησαυροί, τα ελεγχόμενα λεξιλόγια αλλά και τα λεξιλόγια καθιερωμένων όρων, όπως οι όροι της Βιβλιοθήκης του Κογκρέσου (LCSH) και το διεθνές αρχείο καθιερωμένων όρων (VIAF).

Όπως αναφέρθηκε και σε προηγούμενη παράγραφο του παρόντος, τα ελεγχόμενα λεξιλόγια αποτελούν μια προσεκτικά επιλεγμένη αλληλουχία από λέξεις και φράσεις οι οποίες χρησιμοποιούνται τόσο στον τομέα των Ψηφιακών Βιβλιοθηκών όσο και γενικότερα στο πεδίο της Επιστήμης των Πληροφοριών για την επισημείωση τμημάτων πληροφορίας που μπορεί να αποτελούν κομμάτι μιας ιστοσελίδας ή ενός αρχείου, έτσι ώστε τα παραπάνω να μπορούν να ανακτηθούν με έναν εύκολο τρόπο στα πλαίσια μιας αναζήτησης. Τα ελεγχόμενα λεξιλόγια χρησιμοποιούνται ευρέως στους παραπάνω τομείς, καθώς η χρήση τους έχει οδηγήσει στον περιορισμό του προβλήματος της αμφισημίας και της πολυσημίας, όπου η ίδια έννοια μπορεί να αποδίδεται με περισσότερους από έναν τρόπους, ενισχύοντας έτσι τη συνέπεια του περιεχομένου ενός τεκμηρίου/εγγράφου.

Από την άλλη πλευρά οι οντολογίες αποτελούν έναν από τους πυλώνες του Σημασιολογικού Ιστού, καθώς μπορεί να θεωρηθούν ως ένας ειδικός τύπος λεξιλογίου ή μερικές φορές ακόμα και ως μια συλλογή από URIs (Sauermann & Cyganiak, 2008) τα οποία μπορεί να οδηγούν σε κάποια περιγραφή. Οι οντολογίες συνήθως ακολουθούνται από κάποιο έγγραφο γραμμένο σε κάποια γλώσσα περιγραφής οντολογιών όπως η RDFS (Brickley & Guha, 2014) και OWL. Ανάλογα με το πεδίο εφαρμογής έχουν αναπτυχθεί διαφορετικές οντολογίες, όπως η FOAF (Friend Of A Friend) για την περιγραφή ατόμων, το λεξιλόγιο AGROVOC (http://aims.fao.org/standards/agrovoc) για τον τομέα της αγροτικής παραγωγής, ενώ στον τομέα των βιβλιοθηκών χρησιμοποιούνται εκτενέστατα οι οντολογίες DC (Dublin Core), DCterms (Dublin Core terms), BIBO (Bibliographic Ontology) και SKOS (Simple Knowledge Organization System).

Επιπλέον, οι τεχνολογίες του Σημασιολογικού Ιστού παρουσιάζουν ένα αρκετά ενδιαφέρον ποσοστό ενσωμάτωσης στον τομέα των Βιβλιοθηκών και όχι μόνο. Αρκετές βιβλιοθήκες κάνουν διαθέσιμα τα δεδομένα τους με χρήση των παραπάνω τεχνολογιών, όπως το αρχείο καθιερωμένων όρων της Γερμανικής Εθνικής Βιβλιοθήκης (http://www.dnb.de/EN/lds) ή η Βρετανική Βιβλιοθήκη (http://bnb.data.bl.uk/) που διαθέτει τα δεδομένα της με τη μορφή Διασυνδεδεμένων Ανοικτών Δεδομένων ακολουθώντας τους κανόνες που έχουν καθοριστεί από το συγκεκριμένο πρότυπο (Baker et al., 2011˙, Berners-Lee, 2006).

Για όλους του παραπάνω λόγους είναι φανερό πως υπάρχουν πολύ μεγάλες ευκαιρίες από τον συνδυασμό των μεθόδων εξαγωγής φράσεων από κείμενα και των τεχνολογιών του Σημασιολογικού Ιστού στις Ψηφιακές Βιβλιοθήκες. Μια βασική απαίτηση που υπάρχει στο πεδίο των ακαδημαϊκών ψηφιακών βιβλιοθηκών αποτελεί η αποσαφήνιση του συγγραφέα μιας δημοσίευσης (author disambiguation) (Tan et al., 2006). Η διαδικασία αυτή είναι πολύ σημαντική, καθώς συνδέει ένα επιστημονικό έργο με συγκεκριμένους συγγραφείς και αποκτά ιδιαίτερο νόημα όταν υπάρχουν συγγραφείς που φέρουν το ίδιο όνομα. Στο σύστημα που παρουσιάζεται στην εργασία των Tan et al. πραγματοποιείται μια ανάλυση των αποτελεσμάτων από αναζητήσεις που δημιουργούνται αυτόματα. Αποδεικνύεται, επίσης, ότι οι πληροφορίες που υπάρχουν σε ιστοσελίδες που δεν εμφανίζονται στις πρώτες θέσεις των αποτελεσμάτων που επιστρέφονται από τη μηχανή αναζήτησης περιέχουν πιο ακριβείς πληροφορίες σε σχέση με άλλες πιο κοινές ιστοσελίδες, ένα γεγονός που έρχεται σε αντίθεση με ό,τι θα περίμενε κάποιος.

Στην παραπάνω περίπτωση εάν χρησιμοποιηθεί κάποια μέθοδος εξαγωγής φράσεων με τη βοήθεια κάποιου λεξιλογίου για την επισημείωση των εγγράφων αυτών είναι δυνατόν να ξεχωρίσουν οι δημοσιεύσεις δύο συγγραφέων που φέρουν το ίδιο όνομα, καθώς πιθανότατα θα αναφέρονται σε διαφορετικό πεδίο εφαρμογής και άρα η λίστα των φράσεων που θα προκύπτει θα είναι διαφορετική. Ακόμα, με χρήση των τεχνολογιών του Σημασιολογικού Ιστού και χρησιμοποιώντας τις έννοιες των οντολογιών θα μπορούσαν πιθανόν να εξαχθούν οι σχέσεις μεταξύ ενός συγγραφέα και των δημοσιεύσεων του με έναν αυτόματο και ταυτόχρονα αποδοτικό τρόπο.

Μία ακόμα προσέγγιση αποτελεί αυτή που παρουσιάζεται από τους Sugiyama,Kumar, Kan, Tripathi (2010), η προσπάθεια των οποίων επικεντρώνεται στην αυτόματη εξαγωγή και ανάλυση των προτάσεων που χρήζουν αναφοράς σε μια δημοσίευση. Αυτό θα μπορούσε να βοηθήσει στην ταυτοποίηση ενός τμήματος κειμένου σχετικά με το αν χρειάζεται να μπει σ' αυτό μια αναφορά ή όχι. Σ' αυτή την προσέγγιση χρησιμοποιείται ένας ταξινομητής (classifier) που υλοποιεί τεχνικές μηχανικής μάθησης, λαμβάνοντας υπόψη απλά χαρακτηριστικά που υπάρχουν μέσα στο κείμενο, όπως ουσιαστικά αλλά και τις λέξεις που υπάρχουν στην αρχή και το τέλος κάθε πρότασης. Σκοπός αυτού του ταξινομητή είναι να εντοπίσει εάν μια πρόταση χρειάζεται αναφορά ή όχι. Η τεχνική αυτή χρησιμοποιεί μεθόδους επεξεργασίας φυσικής γλώσσας και μπορεί να εφαρμοστεί αποδοτικά σε ένα έξυπνο περιβάλλον συγγραφής, στο οποίο τα αποτελέσματα της μεθόδου αυτής θα μπορούσαν να υποβοηθούν τους συγγραφείς για τα σημεία που χρειάζονται αναφορά. Με χρήση των τεχνικών εξαγωγής φράσεων από κείμενα αλλά και την εύρεση σχετικών με το αντικείμενο του συγγραφέα άρθρων, η συγγραφή θα μπορούσε περαιτέρω να υποβοηθηθεί προτείνοντας στον συγγραφέα σχετικά άρθρα που πιθανόν θα ήθελε να ενσωματώσει στην εργασία του.

Από τα παραπάνω γίνεται φανερό πως υπάρχουν αναρίθμητες ευκαιρίες ενσωμάτωσης των τεχνολογιών Εξαγωγής Φράσεων και του Σημασιολογικού Ιστού στις Ψηφιακές Βιβλιοθήκες τόσο στις ελληνικές όσο και στις διεθνείς. Είναι, επίσης, φανερό ότι προκύπτουν πολλαπλά οφέλη από την ενσωμάτωση των τεχνολογιών αυτών, με τελικούς αποδέκτες συγγραφείς, βιβλιοθηκονόμους αλλά και τελικούς χρήστες.

5. Συμπεράσματα

Στο παρόν άρθρο επιχειρήθηκε μια σύντομη εισαγωγή και επισκόπηση των μεθόδων Εξαγωγής Φράσεων από κείμενα αλλά και της αυξανόμενης ενσωμάτωσης των μεθόδων αυτών σε ένα πλήθος εφαρμογών, όπως για τον εμπλουτισμό μεταδεδομένων των τεκμηρίων, που μπορεί στη συνέχεια να υποβοηθήσουν την πλοήγηση και την αναζήτηση. Η βασικότερη τεχνική που χρησιμοποιείται αφορά μεθόδους μηχανικής μάθησης οι οποίες συνδυάζονται και με άλλες τεχνικές, όπως οντολογίες ή ελεγχόμενα λεξιλόγια, για να πετύχουν το επιθυμητό αποτέλεσμα. Στη συνέχεια αναλύθηκαν τα πολλαπλά οφέλη που προκύπτουν από την ενσωμάτωση τέτοιων μεθόδων σε ποικίλα συστήματα και πεδία εφαρμογής, καθώς και οι δυνατότητες που προσφέρουν οι μέθοδοι αυτές αν εφαρμοστούν στον τομέα των βιβλιοθηκών, παρέχοντας προηγμένες υπηρεσίες στους τελικούς χρήστες.

6. Βιβλιογραφία

Baker, T., Bermès, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., Panzer, M., Schneider, J., Singer, R., Summers, E., Waites, W., Young, J. & Zeng, M. (2011). *Library Linked Data Incubator Group Final Report*. Ανακτήθηκε 15 Ιουλίου, 2015, από http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Barker, K. & N. Cornacchia (2000). Using noun phrase heads to extract document keyphrases. In Proc. of the *13th Canadian Conference on Artificial Intelligence*, pp. 40–52.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American, 284* (5), 28-37.

Berners-Lee, T. (2006). *Linked Data - Design Issues*. Ανακτήθηκε 15 Ιουλίου, 2015, από http://www.w3.org/DesignIssues/LinkedData.html

Brickley, D. & Guha, R.V. (2014). *RDF Schema 1.1*. Ανακτήθηκε 15 Ιουλίου, 2015, από http://www.w3.org/TR/rdf-schema/

Feather, J. & P. Sturges (1996). *International Encyclopedia of Information and Library Science.* London & New York: Routledge.

Han, H., Gilles, C.L., Manavoglu E., Zha, H., Zhang, Z., Fox, E.A. (2003). Automatic Document Extraction using Support Vector Machines, *Proceedings of the 2003 Joint Conference on Digital Libraries.*

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F. & Rudolph, S. (2012). *OWL 2 Web Ontology Language Primer*. Ανακτήθηκε 15 Ιουλίου, 2015, από http://www.w3.org/TR/owl2-primer

Hulth, A. (2004). *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction.* Ph. D. thesis, Computer and Systems Sciences, Stockholm University.

Jones, S. & M. Mahoui (2000). Hierarchical document clustering using automatically extracted keyphrases. In *Proc. of the 3rd International Asian Conference on Digital Libraries*, pp. 113–120.

Kim, S. N., & Baldwin, T. (2009). The Use of Topic Representative Words in Text Categorization University of Melbourne. In *Proceedings of the fourteenth Australasian document computing symposium* (ADCS 2009) (pp. 75-81). Sydney, Australia.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics, 11* (1-2), 11–31.

Medelyan, O. (2005). *Automatic keyphrase indexing with a domain-specific thesaurus.* Master's thesis, Albert-Ludwigs University.

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 296-297). ACM.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14* (3), 130–137.

Sauermann, L. & Cyganiak, R. (2008). *Cool URIs for the Semantic Web*. Ανακτήθηκε 15 Ιουλίου, 2015, από http://www.w3.org/TR/cooluris

Schreiber, G. & Raimond, Y. (2014). *RDF 1.1 Primer*. Ανακτήθηκε 15 Ιουλίου, 2015, από http://www.w3.org/TR/rdf11-primer/

Sugiyama, K., Kumar, T., Kan, M. Y., & Tripathi, R. C. (2010). Identifying citing sentences in research papers using supervised learning. *International Conference In Information Retrieval & Knowledge Management*, (CAMP), 67-72. IEEE.

Tan, Y.F., Kan, M.-Y., Lee D. (2006). Search Engine Driven Author Disambiguation. *Proceedings of the 2006 Joint Conference on Digital Libraries (JCDL).* Chapel Hill, North Carolina, USA.

Turney, P. (1999). *Learning to extract keyphrases from text.* Technical report, National Research Council Canada.

Witten, I. H., G.W. Paynter, E. Frank, C. Gutwin, & C. G. Nevill-Manning (1999). Kea: Practical automatic keyphrase extraction. In *Proc. of the 4th ACM Conference on Digital Libraries (DL 99)*, pp. 254–255. Berkeley, CA: ACM Press.

Zervanou, K., Korkontzelos, I., Van Den Bosch, A., & Ananiadou, S. (2011). Enrichment and structuring of archival description metadata. In P*roceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 44-53. Association for Computational Linguistics.

Zimmermann, A., Lopes, N., Polleres, A., & Straccia, U. (2012). A general framework for representing, reasoning and querying with annotated Semantic Web data. *Web Semantics: Science, Services and Agents on the World Wide Web, 11*, 72-95.

Υ.Δ Ελένη Θ. Γιαννοπούλου

Η Ελένη Γιαννοπούλου αποφοίτησε από το Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας το 2008. Είναι κάτοχος μεταπτυχιακού τίτλου στα «Προηγμένα Συστήματα Πληροφορικής» του Πανεπιστημίου Πειραιώς (2010). Από το 2010 είναι υποψήφια διδάκτωρ στο ΕΜΠ και μέλος της ερευνητικής ομάδας «Επικοινωνιών Πολυμέσων και Τεχνολογιών Διαδικτύου». Τα ερευνητικά της ενδιαφέροντα εστιάζονται στο πεδίο του Σημασιολογικού Ιστού, Ιδρυματικών Αποθετηρίων & Ψηφιακών Βιβλιοθηκών σχετικά με την οργάνωση και την παροχή ψηφιακού περιεχομένου Ανοικτής Πρόσβασης.