Ανάπτυξη πολυγλωσσικού θησαυρού από ετερογενείς πηγές

Μαρία Γαβριηλίδου

ΠΕΡΙΛΗΨΗ

Στην ανακοίνωση παρουσιάζεται ένας τρίγλωσσος θησαυρός όρων ο οποίος υποστηρίζει ένα σύστημα αναζήτησης και ανάκτησης πληροφοριακού υλικού από ετερογενείς πληροφοριακούς πόρους σχετικά με τον πολιτισμό και την ιστορία του Πόντου και της Μικράς Ασίας[1]. Η ανάπτυξη του θησαυρού βασίστηκε στο υλικό δύο φορέων, οι οποίοι έδρασαν ως πάροχοι περιεχομένου αλλά και ως χρήστες. Αναλύεται η διαδικασία κατασκευής του θησαυρού: περιγράφεται το υλικό στο οποίο βασίστηκε ο θησαυρός και οι δύο προσεγγίσεις που ακολουθήθηκαν λόγω των ιδιαιτεροτήτων κάθε συλλογής και παρουσιάζεται ο τρίγλωσσος (ελληνικά, αγγλικά, τουρκικά) θησαυρός που υλοποιήθηκε.

Development of a multi-lingual thesaurus from heterogeneous sources

Maria Gavriilidou

ABSTRACT

This paper presents a trilingual thesaurus which supports an information retrieval system accessing heterogeneous resources concerning the culture and history of Pontus and Asia Minor. The development of the thesaurus was based on the collections of two organisations which acted as content providers and as users. The process of thesaurus construction is analysed: the description of the content on which the thesaurus was based is followed by the presentation of the two approaches adopted due to the idiosyncrasies of each collection and the paper concludes with the description of the trilingual (Greek, English, Turkish) thesaurus.

0 Εισαγωγή

Ο θησαυρός που παρουσιάζεται στην ανακοίνωση αυτή προέκυψε ως η βέλτιστη απάντηση στην ανάγκη για ενιαία προσπέλαση του ετερογενούς υλικού δύο φορέων πολιτιστικού περιεχομένου. Η ετερογένεια οφειλόταν στον αριθμό και τα είδη των τεκμηρίων, στη μορφή

[1] Η ανάπτυξη του θησαυρού έγινε στο πλαίσιο του έργου ΕΗΓ-89 «Κοινοτικός Ιστός του Ελληνισμού της Διασποράς», Πρόγραμμα «Επεξεργασία Εικόνων, Ήχου και Γλώσσας», Μέτρο 3.3, Επιχειρησιακό Πρόγραμμα Κοινωνία της Πληροφορίας.

και στη δομή τους, στην ποιότητα και στο σχήμα ταξινόμησης ή/και τεκμηρίωσης του καθενός. Παρόλες όμως τις διαφορές τους, οι συλλογές υλικού των δύο φορέων αφορούν την ίδια θεματική περιοχή: την ιστορία και τον πολιτισμό του Ελληνισμού της Μικράς Ασίας και του Πόντου. Η λύση που μπορεί να προσφέρει ο θησαυρός είναι η οργάνωση του υλικού με τρόπο κοινό για όλες τις πηγές˙ καταγράφει τις έννοιες του γνωστικού αντικειμένου και αναδεικνύει τις σημασιολογικές σχέσεις που τις συνδέουν. Συνδέοντας τους όρους κάθε πηγής με τις έννοιες του θησαυρού, ουσιαστικά μεταφέρεται η αναζήτηση των τεκμηρίων από το επίπεδο της λέξης στο επίπεδο της έννοιας, με συνέπεια να ελαχιστοποιείται η εξάρτηση από συγκεκριμένους πόρους και την ορολογία που ο καθένας χρησιμοποιεί. Με τον τρόπο αυτό διευκολύνεται η αναζήτηση και ο θησαυρός προσφέρεται για αξιοποίηση από άλλους πόρους περιεχομένου στον ίδιο θεματικό τομέα.

1 Τι είναι ένας θησαυρός

Ο θησαυρός αποτελεί μία δομημένη βάση γνώσης, που καταγράφει τις έννοιες ενός τομέα και τις σχέσεις που ενυπάρχουν ανάμεσα στις έννοιες, και χρησιμοποιείται ως ελεγχόμενο λεξιλόγιο για τις διαδικασίες διαχείρισης κειμενικής πληροφορίας (και κατεξοχήν για ευρετηρίαση / δεικτοδότηση (text indexing) και ανάκτηση κειμένου (text retrieval)). Τεχνικές αναζήτησης που βασίζονται σε απλές αναζητήσεις συμβολοσειρών είναι υπολογιστικά ταχύτατες, ταυτόχρονα όμως και ανεπαρκείς, δεδομένου ότι δεν μπορούν να χειριστούν ιδιότητες της φυσικής γλώσσας όπως:

- η συνωνυμία (π.χ. *δουλεία / σκλαβιά, ενδυμασία / φορεσιά*)
- η ομωνυμία (π.χ. *όργανο* (μουσικό) / *όργανο* (της τάξης)) ή
- η πολυσημία (π.χ. *καλλιέργεια* στη γεωργία και στον πολιτισμό).

Οι σημασιολογικές αυτές σχέσεις καθώς και οι σχέσεις που αντανακλούν εννοιολογικές ιεραρχίες, αναπαρίστανται στους θησαυρούς με τη μορφή δικτύων, στα οποία οι λέξεις/όροι αποτελούν κόμβους που συνδέονται με ονομασμένες σχέσεις. Η χρήση θησαυρών καθιστά την δεικτοδότηση και την αναζήτηση πιο αποτελεσματικές, δεδομένου ότι η προσθήκη εννοιολογικής πληροφορίας αυξάνει την ευφυΐα του συστήματος.

Οι βασικές σχέσεις που καθορίζονται από τα σχετικά πρότυπα (ΕΛΟΤ 1321, ISO 2788, ISO 5964 και ANSI/ NISOZ39.19) είναι:

- **ιεραρχικές**: σχέσεις υπερωνυμίας / υπωνυμίας (ευρύτερος / στενότερος όρος),
- **ισοδυναμίας**: σχέσεις συνωνυμίας,

- **συσχετιστικές**: μη ιεραρχικές σημασιολογικές σχέσεις που ισχύουν ανάμεσα σε όρους, όπως π.χ. ανάμεσα σε ένα αντικείμενο και τις ιδιότητές του.

Με βάση τη δομή του θησαυρού, ένα σύστημα ανάκτησης κειμένου είναι σε θέση να επεκτείνει ή αντίθετα να περιορίσει την ερώτηση του χρήστη στο σύστημα (query expansion / refinement) και να εντοπίσει τα κείμενα που περιλαμβάνουν τον ζητούμενο όρο με μεγαλύτερη ακρίβεια. Οι σχέσεις μεταξύ των όρων του θησαυρού επιτρέπουν, κατά τη διαδικασία αναζήτησης πληροφοριών, την ανάκτηση κειμένων με την μεγαλύτερη δυνατή πληρότητα (recall) και ακρίβεια (precision). Συνεπώς, μέσω της αναζήτησης με έναν όρο, είναι δυνατό να ανακτώνται κείμενα που περιέχουν όχι μόνο τον συγκεκριμένο όρο, αλλά και κείμενα που περιέχουν τους σχετιζόμενους με αυτόν όρους, δηλαδή τους συνώνυμους, τους υπώνυμους, τους υπερώνυμους κτλ. όρους, ώστε να καταστεί πληρέστερη και αποτελεσματικότερη η ανταπόκριση από το σύστημα.

Για την κατάρτιση του συγκεκριμένου θησαυρού μελετήθηκαν διεξοδικά υπάρχοντες θησαυροί σχετικοί με τον πολιτισμό, όπως ενδεικτικά:

- ο Θησαυρός της UNESCO (http://www2.ulcc.ac.uk/unesco/index.htm) είναι ελεγχόμενο λεξιλόγιο οργανωμένο στους τομείς: εκπαίδευση, θετικές επιστήμες, πολιτισμό, κοινωνικές και ανθρωπιστικές επιστήμες, πληροφορία και επικοινωνία, πολιτική, νομικά και οικονομικά και κράτη.
- ο Θησαυρός Τέχνης και Αρχιτεκτονικής (The Art & Architecture Thesaurus – AAT) και ο Θησαυρός Γεωγραφικών Ονομάτων (The Getty Thesaurus of Geographic Names - TGN) του Ινστιτούτου Getty (http://www.getty.edu/research/conducting_research/vocabularies/), οι οποίοι καλύπτουν τους τομείς της τέχνης, της αρχιτεκτονικής και του υλικού πολιτισμού.
- ο Εθνογραφικός Θησαυρός (The Ethnographic Thesaurus Project) (http://et.afsnet.org), της Αμερικανικής Λαογραφικής Εταιρίας και της Βιβλιοθήκης του Κογκρέσου, που εστιάζει στους τομείς της λαογραφίας, της κοινωνικής ανθρωπολογίας, της εθνομουσικολογίας κ.ά.

Εξ αντικειμένου, ωστόσο, ο θησαυρός του Μικρασιατικού και του Ποντιακού Ελληνισμού δεν θα μπορούσε να προκύψει από μετάφραση ή έστω και προσαρμογή των θησαυρών αυτών, δεδομένου ότι, ειδικά στον χώρο της κοινωνικής και πολιτισμικής ανθρωπολογίας, δεν είναι καθόλου αυτονόητο ότι το πλέγμα των εννοιών θα συμπίπτει από το ένα πολιτισμικό και γλωσσικό περιβάλλον στο άλλο. Με το δεδομένο αυτό, δημιουργήθηκε ένας νέος θησαυρός

(με διαδικασία bottom-up), ο οποίος βασίστηκε στο υλικό των φορέων και στην τεκμηρίωσή του, όπου υπήρχε.

2 Η κατασκευή του θησαυρού

2.1 Το υλικό των πολιτιστικών φορέων και η μέθοδος αξιοποίησής του

Η ετερογένεια του υλικού των δύο φορέων υπαγόρευσε διαφορετικές προσεγγίσεις για την άντληση της απαιτούμενης πληροφορίας προκειμένου για την κατάρτιση του θησαυρού. Το υλικό των φορέων παρουσιάζεται αναλυτικά παρακάτω.

2.1.1 Υλικό Επιτροπής Ποντιακών Μελετών (http://www.epm.gr/)

Τα τεκμήρια που απαρτίζουν τη συλλογή της Επιτροπής Ποντιακών Μελετών (ΕΠΜ) προέρχονται από περιοδικά, βιβλία, εφημερίδες, φωτογραφίες και ποικίλο αρχειακό υλικό και αναφέρονται στον εν γένει πολιτισμό του ποντιακού ελληνισμού.

Το υλικό βρίσκεται στο στάδιο της ψηφιοποίησης· διαφέρει, ωστόσο, ο βαθμός ολοκλήρωσης της διαδικασίας για κάθε επιμέρους κατηγορία τεκμηρίων. Η τεκμηρίωση όλου του υλικού, ψηφιοποιημένου και μη, έχει ολοκληρωθεί, και το σχήμα τεκμηρίωσης είναι ενιαίο ως προς τις προδιαγραφές αλλά και προσαρμοσμένο στις ιδιαιτερότητες κάθε κατηγορίας τεκμηρίων.

Όλα τα τεκμήρια συνοδεύονται από πληροφορία για τα θέματα τα οποία πραγματεύεται το καθένα, τα οποία κωδικοποιούνται μέσω λέξεων-κλειδιών. Συμπληρωματικά με τα ανωτέρω πεδία, τα τεκμήρια της ΕΠΜ συνοδεύονται από τον κατάλληλο κωδικό του Ταξινομικού Συστήματος Dewey και την αντίστοιχη περιγραφή (Dewey Decimal Classification, DDC). Στο Σχήμα 1 παρουσιάζονται ενδεικτικά παραδείγματα ορισμένων πεδίων του σχήματος τεκμηρίωσης.

Δεδομένου ότι το υλικό της ΕΠΜ είχε εν μέρει μόνο ψηφιοποιηθεί, αλλά πλήρως τεκμηριωθεί, για την κατάρτιση του θησαυρού αξιοποιήθηκε η υπάρχουσα πληροφορία από το σύστημα τεκμηρίωσης, και όχι το ίδιο το υλικό. Από το σύνολο της πληροφορίας επιλέχθηκαν οι λέξεις-κλειδιά που υπήρχαν στο πεδίο Θέμα και Τόπος σε συνδυασμό με την ταξινόμηση κατά Dewey, και αξιοποιήθηκαν για την κατάρτιση του θεματικού και του γεωγραφικού θησαυρού, αντίστοιχα.

Id	Θέμα	Τόπος	Χρον/γία	Συλλογή	Κωδικός Dewey
3849	Αγρότες	Πόντος	1204-1461	Αρχείον Πόντου	307.7 Specific kinds of communities
8662	Αϊ Κώστας (Βουνό)	Χαλδία		Ποντιακή Εστία	915.65 Geography of East Central Turkey
5525	Ακριτικά δημοτικά τραγούδια			Χρονικά του Πόντου	398.87 Folk songs
10394	Αλέξανδρος Δηλανάς, μητροπολίτης Βεροίας-Ναούσης		1878-1958	Ποντιακή Εστία	270.092 Persons in Christian church history

Σχήμα 1: Ενδεικτικά παραδείγματα πεδίων του σχήματος τεκμηρίωσης της ΕΠΜ

2.1.2 Υλικό Ομοσπονδίας Ποντιακών Σωματείων Νοτίου Ελλάδος (ΟΠΣΝΕ)

Το υλικό της ΟΠΣΝΕ είναι πλήρως ψηφιοποιημένο αλλά όχι τεκμηριωμένο. Περιλαμβάνει φωτογραφικό υλικό, CD, βίντεο και κειμενικό υλικό που αποτελείται από μελέτες σε θέματα ιστορίας και λαογραφίας του Πόντου. Δεδομένου ότι δεν υπάρχει τεκμηρίωση για το υλικό της ΟΠΣΝΕ, ήταν όμως εφικτή η ψηφιακή προσπέλαση στο κειμενικό υλικό, η προσέγγιση που ακολουθήθηκε στην περίπτωση αυτή είναι διαφορετική από την προηγούμενη: βασίζεται στην τεχνολογία εξαγωγής όρων από κείμενα.

Η προσέγγιση του συστήματος εξαγωγής ορολογίας το οποίο χρησιμοποιήθηκε [4], [5] είναι υβριδική: περιλαμβάνει συνδυασμό γλωσσολογικών και στατιστικών μεθόδων. Τα βασικά στάδια του συστήματος εξαγωγής όρων είναι δύο:

(α) εξαγωγή πολυλεκτικών όρων με βάση γλωσσικούς κανόνες

Οι κανόνες που χρησιμοποιούνται περιγράφουν τις συντακτικές δομές με τις οποίες εμφανίζονται οι πολυλεκτικοί όροι (π.χ. ΕΠΙΘ + ΟΥΣ, ΟΥΣ + ΟΥΣγεν., ΟΥΣ + ΠΡΟΘ + ΟΥΣ κτλ.). Η αδυναμία της γραμματικής έγκειται στο ότι εφαρμόζει τους κανόνες της χωρίς διάκριση, περιγράφοντας την ικανή αλλά όχι και αναγκαία συνθήκη για να είναι όρος μια ακολουθία λέξεων. Επιπλέον, μπορεί να εντοπίσει μόνο πολυλεκτικούς όρους, δεδομένου ότι μόνο σε αυτούς μπορεί να αποδοθεί συντακτική δομή. Ο απώτερος στόχος μιας γραμματικής όρων είναι ο εντοπισμός “υποψήφιων όρων”, η επικύρωση των οποίων χρειάζεται να γίνει από ένα εργαλείο διαφορετικής φύσης ή από ειδικούς.

(β) εξαγωγή μονολεκτικών όρων και στατιστική επιβεβαίωση πολυλεκτικών όρων

Η στατιστική προσέγγιση στηρίζεται στην υπόθεση ότι οι όροι, ως λέξεις ή φράσεις που είναι

χαρακτηριστικές του θεματικού πεδίου του κειμένου, τείνουν να έχουν υψηλή συχνότητα στο κείμενο. Στα μειονεκτήματα της στατιστικής προσέγγισης καταγράφεται η αδυναμία να εξάγει όρους που δεν ικανοποιούν τα στατιστικά κριτήρια· ωστόσο, είναι πιθανό έγκυροι όροι να εμφανίζονται μόνο μία ή λίγες φορές στο κείμενο και γι' αυτό να μην εντοπίζονται από τη στατιστική μέθοδο.

Με την εφαρμογή του συστήματος προέκυψε ένας κατάλογος μονολεκτικών και πολυλεκτικών υποψήφιων όρων· ακολούθησε επεξεργασία των υποψηφίων όρων, με στόχο την επιλογή των όρων που θα αποτελούσαν τον τελικό κατάλογο όρων της ΟΠΣΝΕ (π.χ. επικύρωση του υποψήφιου όρου *μεικτός χορός*, αλλά απόρριψη του υποψήφιου όρου *διαφορετική μουσική*), λημματοποίηση (π.χ. *αστικά κέντρα → αστικό κέντρο*) και διορθώσεις (π.χ. *τουρκική ονομασία* και *τούρκικη ονομασία* ενοποιήθηκαν ως *τουρκική ονομασία*).

2.2 Θησαυροί όρων

Από την σύγκριση και ενοποίηση των καταλόγων όρων των δύο φορέων δημιουργήθηκαν δύο θησαυροί: ο γεωγραφικός και ο θεματικός θησαυρός.

2.2.1 Γεωγραφικός θησαυρός

Ο γεωγραφικός θησαυρός περιλαμβάνει κατηγορίες κατάλληλες για την ταξινόμηση των τοπωνυμίων που απαντούν στο κειμενικό υλικό των φορέων. Βασικό πρόβλημα για τη δόμηση των κατηγοριών του γεωγραφικού θησαυρού αποτέλεσαν οι διαφορετικές γεωγραφικές διαιρέσεις των περιοχών στις οποίες αναφέρεται το υλικό των φορέων κατά τη διάρκεια της ιστορίας τους. Το πρόβλημα οφείλεται στο γεγονός ότι το υλικό αναφέρεται τόσο σε σύγχρονες γεωπολιτικές οντότητες (π.χ. στα σύγχρονα κράτη, πόλεις κτλ.), όσο και σε γεωγραφικές περιοχές οι οποίες σε παλαιότερες εποχές ανήκαν σε διαφορετικά κράτη με διαφορετικά σύνορα απο τα σημερινά, αλλά και, επιπροσθέτως, στο ότι συνυπήρχαν διαφορετικές οργανώσεις των στοιχείων αυτών εντός της ίδιας ιστορικής περιόδου. Η ύπαρξη διαφορετικής διοικητικής και εκκλησιαστικής διαίρεσης των περιοχών της Μικρασίας και του Πόντου είχε ως αποτέλεσμα συχνά να εντάσσεται ένα τοπωνύμιο σε μία διοικητική περιφέρεια αλλά σε διαφορετική εκκλησιαστική – με τις εκκλησιαστικές περιφέρειες, ωστόσο, να έχουν και διοικητικό χαρακτήρα στη δομή της Οθωμανικής Αυτοκρατορίας.

Πρόσθετο θεωρητικό πρόβλημα που συνδέεται με το γεωγραφικό θησαυρό είναι το ότι ορισμένες γεωγραφικές / γεωπολιτικές οντότητες εμφανίζονται με διάφορα ονόματα στη διάρκεια της ιστορικής τους ύπαρξης. Αυτό οφείλεται είτε στην εξέλιξη των τοπωνυμίων ενδογλωσσικά σύμφωνα με τη γενική εξέλιξη της γλώσσας, είτε στην εξωγλωσσική (δηλαδή με διοικητική ρύθμιση) αλλαγή του τοπωνυμίου, είτε στη διαφορετική ονομασία ανάλογα με

τη γλώσσα (Ελληνική / Τουρκική), είτε, τέλος, στο γεγονός ότι η συνύπαρξη των γλωσσών επιφέρει τροποποιήσεις των τοπωνυμίων. Για παράδειγμα, η αρχαία *Αμισός* έγινε *Σαμψούς / Σαμψούντα* μέσω του τουρκικού *Samsun.* Πράγμα που δεν συνέβη με την *Τραπεζούντα* η οποία διατηρεί το όνομά της και δεν επηρεάζεται από το τουρκικό *Trabzon*. Οι διαφορετικές Ελληνικές ονομασίες αντιμετωπίστηκαν ως συνώνυμοι όροι, ενώ οι αντίστοιχες Τουρκικές ονομασίες κωδικοποιήθηκαν στο πεδίο του μεταφραστικού αντιστοίχου, εφόσον, βέβαια, η ταυτοποίησή τους ήταν δυνατή.

Τέλος, πολύ σημαντικό πρόβλημα ταξινόμησης των τοπωνυμίων προέκυψε από το γεγονός ότι το ίδιο το υλικό πολλές φορές δεν έδινε στοιχεία για τη θέση κάποιου τοπωνυμίου ή την ακριβή ένταξή του σε κάποια ευρύτερη διοικητική δομή (νομό / περιφέρεια / τμήμα κτλ.).

Για την επίλυση του θέματος της δόμησης του θησαυρού υιοθετήθηκε ένα λιγότερο αυστηρό σχήμα, το οποίο επιτρέπει την εμφάνιση των ταξινομικών κατηγοριών του θησαυρού όπου και όπως αυτές μπορούσαν να εφαρμοστούν σύμφωνα με την πληροφορία του υλικού.

Έτσι, διαμορφώθηκαν τα εξής υπο-σχήματα του γεωγραφικού θησαυρού ως εξής:

➢ ήπειροι • χώρες o νομοί ▪ πόλεις ▪ χωριά	➢ γεωγραφικές περιοχές • εκκλησιαστικές επαρχίες o περιφέρειες ▪ τμήματα ✧ πόλεις ♦ δήμοι - συνοικίες (μαχαλάδες)
o περιοχές ▪ πόλεις ✧ δήμοι ♦ συνοικίες (μαχαλάδες)	o επαρχίες ▪ πόλεις ✧ δήμοι ♦ συνοικίες (μαχαλάδες)

Έτσι, ο κόμβος «πόλεις» εμφανίζεται ως υπώνυμος άλλοτε (ανάλογα με το ιεραρχικό υπο-σχήμα) του κόμβου «περιοχές», και άλλοτε των κόμβων «νομοί», «τμήματα» ή «επαρχίες». Για παράδειγμα, το τοπωνύμιο *Κασταμονή* εντάσσεται στο θησαυρό ως εξής:

- Εκκλησιαστική επαρχία Νεοκαισαρείας
    - Τμήμα Κασταμονής
        - Κασταμονή

Το γεγονός ότι δεν εμφανίζονται όλα τα επίπεδα ταξινόμησης που προβλέπει ο θησαυρός οφείλεται σε ανεπαρκή στοιχεία από το ίδιο το υλικό· σε περίπτωση που η προσθήκη νέων τεκμηρίων στη συλλογή δώσει την απαιτούμενη πληροφορία, η σύνδεση του τοπωνυμίου με τους σχετικούς κόμβους του θησαυρού θα είναι εφικτή.

2.2.2 Θεματικός θησαυρός

Ο θεματικός θησαυρός οργανώθηκε σε 21 θεματικά πεδία, ο καθορισμός των οποίων καθοδηγήθηκε από το ίδιο το υλικό των φορέων και συμπληρώθηκε από τους υπάρχοντες διεθνώς θησαυρούς, περισσότερο δε από τον Εθνογραφικό Θησαυρό. Τα θεματικά πεδία είναι τα εξής:

1. Γλώσσα
2. Εκπαίδευση
3. Ένοπλες δυνάμεις
4. Έρευνα / θεωρία / μεθοδολογία
5. Ιστορία
6. Καθημερινή ζωή / Έθιμα
7. Κοινωνία
8. Μετακίνηση και εγκατάσταση
9. Μεταφορές
10. Νόμος & Διακυβέρνηση
11. Όντα
12. Παραγωγή
13. Παραγωγή Λόγου & Λογοτεχνία
14. Πίστη / Λατρεία
15. Τεκμηρίωση
16. Τέχνες
17. Υγεία
18. Υλικός πολιτισμός
19. Φιλοσοφία
20. Φύση / Κλίμα
21. Χώρος / Τόπος

Κάθε θεματικό πεδίο οργανώθηκε έτσι ώστε η δομή του να αναπαριστά τους όρους του πεδίου και τις μεταξύ τους σχέσεις. Τα 21 θεματικά πεδία περιλαμβάνουν συνολικά 628 έννοιες, με τις οποίες συνδέονται οι όροι των δύο φορέων.

2.2.3 Κωδικοποίηση των θησαυρών

Για την ανάπτυξη των θησαυρών χρησιμοποιήθηκε η εφαρμογή «Εξερευνητής Θησαυρών»[2]. Δημιουργήθηκαν δύο βάσεις δεδομένων, μία για κάθε θησαυρό (γεωγραφικό και θεματικό). Ο γεωγραφικός θησαυρός περιλαμβάνει 580 όρους και ο θεματικός 2.565 όρους, οι οποίοι προέκυψαν από το υλικό των πηγών. Οι τελικοί κατάλογοι όρων, όπως προέκυψαν μετά την επεξεργασία, «φορτώθηκαν» στο εργαλείο αυτόματα και η πληροφορία που συνοδεύει κάθε όρο κωδικοποιήθηκε με το χέρι. Η ταυτότητα κάθε όρου περιλαμβάνει:

- πληροφορία για το θεματικό πεδίο στο οποίο ανήκει,
- τον κωδικό Dewey (ο οποίος διατηρήθηκε, όπου υπήρχε),
- τη μετάφρασή του στα αγγλικά και στα τουρκικά (οι μεταφράσεις έγιναν από μεταφραστές),
- τις κατάλληλες συνδέσεις με άλλους όρους.

Μέσω των σχέσεων υπερωνυμίας, συνωνυμίας και σχετικότητας, κάθε όρος εντάσσεται ως κόμβος στο θησαυρό και συνδέεται με τους άλλους όρους. Εκτός από τη διασύνδεση των όρων εντός του ίδιου θεματικού πεδίου, το εργαλείο επιτρέπει επιπλέον και τη διαθεματική σύνδεση μεταξύ όρων: όρος από ένα θεματικό πεδίο μπορεί να συνδέεται με συσχετιστική

[2] Η εφαρμογή «Εξερευνητής Θησαυρών» αναπτύχθηκε από την Cognitron ΕΠΕ, τεχνοβλαστό του ΙΕΛ.

σχέση με όρο από άλλο πεδίο. Η σχέση αυτή μπορεί να συνδέει και σημασίες της ίδιας λέξης: για παράδειγμα ο όρος *πρεσβείες* που ανήκει στο θεματικό πεδίο *Νόμος & Διακυβέρνηση* ως υπώνυμος όρος του όρου *Διπλωματία* συνδέεται με τον όρο *πρεσβείες* που είναι υπώνυμος του όρου *Δημόσια κτίρια* και ανήκει στο θεματικό πεδίο *Υλικός πολιτισμός*.

Ο θησαυρός διευκολύνει την ομοιογενή και με ενιαίο τρόπο πρόσβαση στους πληροφοριακούς πόρους των δύο φορέων, χωρίς να είναι απαραίτητη η γνώση των συγκεκριμένων όρων του κάθε φορέα, δεδομένου ότι αυτοί είναι συνδεδεμένοι με τον σχετικό κόμβο στην ιεραρχία του θησαυρού. Η σύνδεση και άλλων πληροφοριακών πόρων στον ίδιο θησαυρό είναι δυνατή εφόσον ακολουθηθεί η αντίστοιχη διαδικασία - όποια από τις δύο προσεγγίσεις προσφέρεται (αξιοποίηση της τεκμηρίωσης ή εξαγωγή όρων από το υλικό), ανάλογα με τη συλλογή.

3 Βιβλιογραφία

[1] ANSI/ NISOZ39.19 ANSI/NISOZ39.19-2005 *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.* Bethesda, Maryland: NISO Press, 2005.

[2] Dewey Decimal Classification (DDC) http://www.oclc.org/dewey/.

[3] Georgantopoulos, B, Piperidis, S., Eliciting Terminological Knowledge for Information Extraction Applications, in Tzafestas, S. (Ed.) *Advances in Intelligent Systems: Concepts, Tools and Applications,* Kluwer Academic Publishers, Microprocessor-based and Intelligent Systems Engineering Series, 1999, Vol. 21.

[4] Georgantopoulos, B., Piperidis, S., A Hybrid Technique for Automatic Term Extraction, *Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications - ACIDCA'2000*, Monastir, Tunisia, 22-24 March 2000, pp. 124-128.

[5] ISO 2788:1986(2nd): Guidelines for the establishment and development of monolingual thesauri.

[6] ISO 5964:1985(1st): Guidelines for the establishment and development of multilingual thesauri.

[7] ΕΛΟΤ 1321:1993: Τεκμηρίωση – Κατευθυντήριες οδηγίες για τη συγκρότηση και ανάπτυξη μονόγλωσσων θησαυρών.

Γαβριηλίδου Μαρία

Γλωσσολόγος

Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 & Επιδαύρου, 151 25 Μαρούσι
Τηλ. 210 6875441 Τηλ/πο 210 6854270
Ηλ. δ/νση maria@ilsp.gr