

Πρόγραμμα Μεταπτυχιακών Σπουδών στην Επιστήμη της Πληροφορίας
Διοίκηση και Οργάνωση Βιβλιοθηκών με έμφαση στις νέες Τεχνολογίες της Πληροφορίας

ΜΑΘΗΜΑ: «ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ»
Διδάσκων : Σαράντος Καπιδάκης

ΕΡΓΑΣΙΑ:

Το πρόβλημα της αποσαφήνισης γεωγραφικών ονομάτων σε μια ψηφιακή βιβλιοθήκη ιστορικού περιεχομένου: το παράδειγμα της Perseus Digital Library

ΦΟΙΤΗΤΡΙΑ:

Θεοδώρα Τσώλη

Αθήνα, Φεβρουάριος 2004

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

σελ.

1.	Πρόλογος	3
2	Εισαγωγή	3
3.	Perseus Digital Library (PDL)	4
3.1	Τι είναι	4
3.2	Τι περιλαμβάνει	4
3.3	Στόχοι	4
3.4	Περιγραφή	5
3.4.1	Ψηφιακά τεκμήρια	5
3.4.2	Επεξεργασία τεκμηρίων	5
3.4.3	Εργαλεία αναζήτησης	5
4	Το ζήτημα της αποσαφήνισης τοπωνυμίων στα πλαίσια της PDL	6
4.1	Το σύστημα γεωγραφικών δεδομένων της PDL	6
4.2	Περιγραφή του προβλήματος της γλωσσικής αμφισημίας και της αποσαφήνισης τοπωνυμίων	7
4.3	Περιγραφή του προβλήματος και σχετικές εργασίες	8
5	Η διαδικασία αποσαφήνισης	9
5.1	Η δημιουργία του γεωγραφικού λεξικού	9
5.2	Κατηγοριοποίηση κα ταυτοποίηση ονομάτων	10
5.3	Αποσαφήνιση πιθανών τοπωνυμίων	10
5.4	Τελική αποσαφήνιση	11
6	Αξιολόγηση του συστήματος	12
7	Περιορισμοί του συστήματος και μελλοντικοί στόχοι	13
8	Σχετικές εργασίες	14
8.1	Alexandria Digital Library (ADL)	14
8.2	GIPSY	14
8.3	Nominator	15
8.4	Nymble	15
9	Η χρησιμότητα του συστήματος και πιθανές εφαρμογές στον ελληνικό χώρο. Ερωτήματα και προβληματισμοί	16
10	Βιβλιογραφία	18
11	Παραρτήματα	
11.1	Παράρτημα Α: Λεξιλόγιο όρων αγγλικής και αποδόσεις στα ελληνικά	20
11.2	Παράρτημα Β: Παράδειγμα αναζήτησης και αποτύπωσης χωρικής πληροφορίας στην Perseus Digital Library	21
11.3	Παράρτημα Γ: Παραδείγματα εγγραφών ψηφιακών γεωγραφικών λεξικών	24
11.4	Παράρτημα Δ: Άρθρο αναφοράς	25

1. Πρόλογος

Οι διεπαφές γεωγραφικών δεδομένων αποτελούν φυσικές, κλιμακούμενες οπτικοποιήσεις των δεδομένων ψηφιακών βιβλιοθηκών, ωστόσο, η μεγάλη ποικιλία αυτών των δεδομένων και των χαρακτηριστικών τους παρουσιάζει ορισμένα συγκεκριμένα προβλήματα που αφορούν στην ταυτοποίηση και αποσαφήνιση τοπωνυμίων.

Σε μια διεπαφή τέτοιου τύπου οι βιβλιοθηκονόμοι πρέπει να ταυτοποιήσουν και να αποσαφηνίσουν τα γεωγραφικά ονόματα και να τα συνδέσουν με πληροφορίες για την θέση τους στο χώρο, τον τύπο τους (π.χ. ποτάμι, βουνό, κ.λ.π.), τον πληθυσμό, τη σχέση τους με άλλες περιοχές, κτλ. πράγμα που δεν μπορεί να γίνει για μεγάλο όγκο κειμένων.

Παρακάτω θα παρουσιαστεί το σύστημα αποσαφήνισης γεωγραφικών ονομάτων που δημιουργήθηκε και χρησιμοποιείται στα πλαίσια της ψηφιακής βιβλιοθήκης Perseus καθώς και η αξιολόγησή του όπως παρουσιάστηκε σε άρθρο των D. Smith και G. Crane εκ των βασικών δημιουργών της ψηφιακής βιβλιοθήκης Perseus (Smith, 2001).

2. Εισαγωγή

Οι γεωγραφικές διεπαφές παρέχουν φυσικές, διαβαθμισμένες οπτικοποιήσεις σε πολλές συλλογές ψηφιακών βιβλιοθηκών. Αν και οι οντολογίες συγκεκριμένων πεδίων ορισμού (domain-specific ontologies) ή η αυτόματη συσταδοποίηση (clustering) από τεκμήρια δίνουν παραγωγικά εργαλεία φυλλομέτρησης, σε πολλές περιπτώσεις, οι χάρτες (αλλά και οι χρονολογικές σειρές) είναι εργαλεία που μπορούν να παρουσιάσουν ένα μεγάλο εύρος πληροφοριών με πιο οικεία μορφή. Όταν τα περιεχόμενα μιας ψηφιακής βιβλιοθήκης περιέχουν γεωγραφικές αναφορές ο χρήστης μπορεί να αντιληφθεί το εύρος και τα σημεία που εστιάζει το ενδιαφέρον της μια συλλογή ή ένα τεκμήριο, μπορεί να αποτυπώσει τις τοποθεσίες που αναφέρονται σε ένα κείμενο ή να βρει πληροφορίες για τα μέρη που αναφέρονται σε ένα χάρτη ή σε μια περιοχή. Στο πρόγραμμα Perseus ο στόχος ήταν η παρουσίαση με αυτόν το τρόπο ιστορικών δεδομένων από την αρχαία Ελλάδα έως την Αμερική του 19^{ου} αι. και η δημιουργία μια πλούσιας ψηφιακής βιβλιοθήκης που είναι διαθέσιμη και στο Διαδίκτυο (<http://www.perseus.tufts.edu>)

Προκειμένου να εκμεταλλευτούν τα οφέλη των γεωγραφικών διεπαφών οι βιβλιοθηκονόμοι θα πρέπει να ταυτοποιήσουν τα γεωγραφικά ονόματα και να τα συνδέσουν με πληροφορίες για τη θέση τους στο χώρο, με τον τύπο τους (π.χ. ποταμός, βουνό, κατοικημένη περιοχή, κτλ.) και με άλλες χρήσιμες πληροφορίες όπως πληθυσμός σε διάφορες εποχές ή σχέση με άλλα μέρη. Για τεκμήρια που είναι πολύ σημαντικά για την ερευνητική εργασία ίσως να αξίζει να ξοδέψει κάποιος άνθρωπος την προσπάθεια και το χρόνο που απαιτείται να προσδιορίσει και να αποσαφηνίσει τα τοπωνύμια. Ωστόσο, αυτού του είδους η χειρωνακτική εργασία είναι αδύνατη για πολύ μεγάλες συλλογές. Ακόμα και αν κάποιος ξόδευε 10 δευτερόλεπτα για κάθε τοπωνύμιο θα χρειαζόνταν 28.000 εργατοώρες για να ελεγχθούν τα περισσότερα από 1 εκ. τοπωνύμια της ψηφιακής βιβλιοθήκης Perseus. Αυτό σημαίνει ότι υπάρχει ανάγκη για αυτόματες ή τουλάχιστο μηχανικά υποβοηθούμενες μεθόδους κατασκευής ψηφιακών βιβλιοθηκών γεωγραφικών αναφορών.

3. Perseus Digital Library (PDL)

3.1 Τι είναι.

Η PDL είναι μια ετερογενής συλλογή κειμένων και εικόνων που αναφέρονται στην αρχαϊκή και κλασσική Ελλάδα (ιστορία, τέχνη, αρχαιολογία), στην πρώιμη Ρωμαϊκή αυτοκρατορία και την σύγχρονη Ευρώπη και Αμερική (ως το 19^ο αιώνα). Τα κείμενα συγκροτούνται σε ένα ενιαίο σύνολο με την χρήση εργαλείων μορφολογικής ανάλυσης, με λεξικά, και εκλεπτυσμένα εργαλεία αναζήτησης που επιτρέπουν στους χρήστες να βρίσκουν για παράδειγμα όλες τις εμφανίσεις ενός συγκεκριμένου λεξιλογικού τύπου (Rydborg-Cox, 2000).

Ξεκίνησε το 1987 στο Harvard ενώ τώρα ο κεντρικός server είναι εγκατεστημένος στο Tufts University (Crane, 1996).

3.2 Τι περιλαμβάνει.

- 9 εκ. λέξεις στην ελληνική και 4 εκ. λέξεις στη λατινική γλώσσα με τις μεταφράσεις τους στα αγγλικά
- 55 εκ. αγγλικές λέξεις
- 3 λεξικά ελληνικών και λατινικών
- χιλιάδες εγγραφές μουσειακών εκθεμάτων
- 5.000 αρχαία τοπωνύμια που αποτυπώνονται σε περισσότερους από 1.000 χάρτες
- 30.000 έγχρωμες φωτογραφίες,
- άρθρα από εγκυκλοπαίδειες και άλλες πηγές
- 375.000 συνδέσμους που αλληλοσυνδέουν αυτούς τους πόρους
- εκατομμύρια αυτόματα παραγόμενους συνδέσμους (Smith, 2000a), (Mahoney, et al. 2000).

3.3 Στόχοι.

- Αρχικά ήταν η δημιουργία μιας συμπαγούς βάσης βιβλιοθηκών υπερκείμενων πάνω σε ένα θέμα (την αρχαία ελληνική φιλολογία). (Crane, 1998)
- Κατόπιν η ψηφιακή βιβλιοθήκη έθεσε ως στόχο την επέκταση από το χώρο της κλασσικής αρχαιότητας σε ένα μεγάλο εύρος θεμάτων και επιστημών. Τα ερευνητικά προγράμματα της PDL επιχειρούν έτσι την ομογενοποίηση μουσειακών εκθεμάτων, αρχαιολογικών τόπων, τοποθεσιών, πρωτογενών και δευτερογενών κειμένων ώστε να επιτευχθεί η αλληλοσυσχέτιση κειμένων και αντικειμένων στο χώρο και το χρόνο. (Smith, 2000a)
- Ο γενικότερος στόχος είναι η ανάπτυξη των ψηφιακών βιβλιοθηκών για το σύνολο των ανθρωπιστικών επιστημών και την παραγωγή εργαλείων που η διαχείρισή τους γίνεται και από ψηφιακά προγράμματα και από τους ανθρώπους – ερευνητές. (Smith, 2000a).

3.4 Περιγραφή.

3.4.1 Ψηφιακά τεκμήρια.

Η πλειονηφία των ψηφιακών τεκμηρίων κωδικοποιείται και επισημαίνεται με τη γλώσσα TEI (Text Encoding Initiative). Ωστόσο, προκειμένου να περιοριστεί το πρόβλημα των διαφόρων εξειδικεύσεων της XML που χρησιμοποιούνται πια ευρέως από διαφορετικές επιστημονικές κοινότητες, η PDL επιτρέπει τη μερική απεικόνιση στοιχείων διαφόρων δηλώσεων τύπων εγγράφων (DTD's) ώστε να είναι δυνατό να ευρετηριάζεται, να αναλύεται και να παρουσιάζεται από το σύστημα μεγάλος αριθμός τεκμηρίων. (Rydberg-Cox, 2000).

3.4.2 Επεξεργασία τεκμηρίων.

Η Ψηφιακή Βιβλιοθήκη Perseus έχει αναπτύξει ένα καθολικό σύνολο εργαλείων για τη διαχείριση των τεκμηρίων σε XML αλλά και σε διάφορες δηλώσεις τύπων εγγράφων (DTD's) προκειμένου να εξάγει δομικά και περιγραφικά μεταδεδομένα από αυτά τα τεκμήρια μετά από κάθε απαίτηση (Smith et al. 2000b).

Παράλληλα αυτό το σύνολο εργαλείων υποστηρίζει τις ανάγκες γλωσσικής επεξεργασίας και παρουσίασης των τεκμηρίων. Ο αφαιρετικός χαρακτήρας των δομικών και περιγραφικών μεταδεδομένων επιτρέπει την ανάπτυξη διαφόρων εργαλείων γλωσσολογικής ανάλυσης, διαχείρισης γνώσης και ανάκτησης πληροφοριών. Τα σχετικά εργαλεία αυτήν την περίοδο περιλαμβάνουν την δημιουργία υπονοούμενων cross-references, τη δημιουργία υπερκειμένων, την παραγωγή των χαρτών με ένα GIS (γεωγραφικό σύστημα πληροφοριών), την ανακάλυψη προτύπων παράλληλης εμφάνισης λέξεων, τη μορφολογική ανάλυση λέξεων κ.ά.

Ειδικότερα, όσων αφορά την αυτόματη παραγωγή χωροχρονικών απεικονίσεων το σύστημα σαρώνει αυτόματα όλα τα κείμενα για την ανεύρεση των τοπωνυμίων και των ημερομηνιών. Κατόπιν αυτά τα δεδομένα συνδέονται με ένα γεωγραφικό σύστημα πληροφοριών (GIS) που επιτρέπει στους χρήστες να παράγουν χρονολόγια ή χάρτες με βάση τις ημερομηνίες και τα τοπωνύμια που αναφέρονται είτε σε ολόκληρο το τεκμήριο είτε σε ένα τμήμα του. Το πόσο ακριβής θα είναι ο χάρτης καθορίζεται από το ποσοστό αποσαφήνισης των τοπωνυμίων. Με τη σειρά του ο χάρτης είναι πύλη προς τη συλλογή αφού ο χρήστης μπορεί να κινηθεί από ένα σημείο του χάρτη προς το τεκμήριο που αναφέρεται σε αυτό. (Rydberg-Cox, 2000).

3.4.3 Εργαλεία αναζήτησης.

- Αναζήτηση λέξεων ή φράσεων σε ελληνικά και λατινικά κείμενα.
- Lookup Tool: αναζήτηση λέξεων-κλειδιών οπουδήποτε (κείμενο, βάση δεδομένων φωτογραφιών, άτλαντας). Τα αποτελέσματα εμφανίζονται ομαδοποιημένα κατά είδος (π.χ. κείμενα, παραπομπές, φωτογραφίες, γλυπτά, τοποθεσία στο χάρτη, κτλ.).
- Αυτόματη, ολοκληρωμένη αναζήτηση: είναι οι αναζητήσεις που δεν έχει ζητήσει ο χρήστης. Δηλαδή, όταν ο χρήστης βλέπει ένα κείμενο, γίνεται ταυτόχρονη εξαγωγή όλης της διαθέσιμης σχετικής πληροφορίας με τη μορφή συνδέσμων στο κείμενο. Οι παραπομπές στην PDL δημιουργούν συνδέσμους διπλής κατεύθυνσης: τα κείμενα «ακούν» τα κείμενα που τα αναφέρουν. (Mahoney, 2000).

4. Το ζήτημα της αποσαφήνισης τοπωνυμίων στα πλαίσια της PDL

4.1 Το σύστημα γεωγραφικών δεδομένων της PDL

Προκειμένου να έχουμε πρόσβαση στο υλικό μιας ψηφιακής βιβλιοθήκης αυτού του είδους απαιτείται η δημιουργία ενός συστήματος πληροφοριών ώστε να είναι δυνατή η ανάκτηση πληροφοριών μέσω ερωτημάτων χωρικού τύπου και περιγραφών γεωγραφικών θέσεων.

Οι τύποι αυτών των ερωτημάτων αυτών είναι: «τι» και «που» δηλαδή «τι υλικό (κείμενα, εικόνες, κτλ.) σχετίζεται με την X τοποθεσία και υπάρχουν στην ψηφιακή βιβλιοθήκη;» και «που βρίσκεται η X τοποθεσία;». Το εργαλείο που συνδέει την πληροφορία με το ερώτημα είναι σχήμα μεταδεδομένων ή ένα ψηφιακό γεωγραφικό λεξικό*. Ένα παράδειγμα αρχιτεκτονικής ψηφιακής βιβλιοθήκης τέτοιου τύπου είναι η Alexandria Digital Library (ADL) (βλ. σελ 14). Η PDL εφαρμόζει ένα παρόμοιο σύστημα χειρισμού γεωγραφικών πληροφοριών προσαρμοσμένο όμως στις ιδιαιτερότητες του υλικού της (υλικό ιστορικής φύσεως όπως πάπυροι, επιγραφές, εικόνες, κτλ.).

Τα δεδομένα γεωγραφικών αναφορών που αλληλεπιδρούν με τα άλλα συστατικά της PDL εξαρτώνται λοιπόν από τη δημιουργία ενός απλού σχήματος μεταδεδομένων. Το αρχικό σχήμα της PDL περιλαμβάνει το όνομα ενός τόπου, την θέση του στο χώρο (σημείο ή συντεταγμένες) και τον τύπο του. Ενώ αυτά τα χαρακτηριστικά επιτρέπουν στο σύστημα να αρχίσει να κάνει τη σύνδεση ανάμεσα στις συντεταγμένες και τις τοποθεσίες η φύση της ίδιας της ιστορικής πληροφορίας επηρέασε τον τρόπο δημιουργίας του σχήματος. Η γεωγραφική ταυτοποίηση ενός αντικείμενου εξαρτάται απόλυτα από τα διαθέσιμα στοιχεία για τη θέση του στο χώρο. Αν αυτά υπάρχουν μπορούν να προστεθούν στο σχήμα μεταδεδομένων έτσι ώστε το αντικείμενο να έχει και γεωγραφική αναφορά. Συχνά όμως συμβαίνει να μην υπάρχουν πηγές τέτοιων στοιχείων για υλικό της αρχαιότητας ή του Μεσαίωνα. Έτσι, ή το ίδιο το αντικείμενο αποτελεί τέτοια πηγή ή το αντικείμενο αποτελεί τη μοναδική αναφορά για ένα χαρακτηριστικό ή τοποθεσία που πια δεν υπάρχει. Ενώ στην ψηφιακή βιβλιοθήκη Alexandria οι διαθέσιμες πηγές μεταδεδομένων που χρησιμοποιήθηκαν ήταν οι: GNIS (Geographic Names Information System) και GNPS (Geographic Names Processing System) στην περίπτωση της PDL αυτές δεν ήταν επαρκείς. Στην πραγματικότητα οι κύριες πηγές δεδομένων ήταν οι υπονοούμενες αναφορές γεωγραφικού χαρακτήρα της ίδιας της βάσης των κειμένων καθώς και οι χάρτες, οι έρευνες, οι αναφορές, κτλ.

* Ένα ψηφιακό γεωγραφικό λεξικό είναι μια λίστα καθιερωμένων γεωγραφικών ονομάτων που αποτελείται μεταξύ άλλων και από: α) ένα όνομα (ή πολλά ονόματα) για ένα τόπο, β) την τοποθεσία του (γεωγραφικές συντεταγμένες), γ) την κατηγορία του τόπου (με βάση ένα σχήμα κατηγοριών που περιγράφουν γεωγραφικά χαρακτηριστικά) (π.χ. Goleta Beach [park], -119.83, 34.42). Βλ. Παράρτημα Γ με τις εγγραφές για την Αθήνα (Ελλάδα) του Getty Thesaurus of Geographic Names (Εικ. 1) και την πόλη Athens, Georgia από το GNIS (Geographic Names Information System) (Εικ. 2)

Προκειμένου να μπορέσουν να «τοποθετηθούν» τα ιστορικά αντικείμενα στο χώρο η PDL ανέπτυξε μια λίστα καθιερωμένων γεωγραφικών τοποθεσιών από πηγές της συλλογής όπως οι: Princeton Encyclopedia of Classical Sites και η Ιστορία του Πελοποννησιακού Πολέμου του Θουκυδίδη οι οποίες συμπλήρωναν λεξικά όπως τα Getty Thesaurus of Geographic Names και Cruchley's gazetteer of London. Αυτός ο καθιερωμένος κατάλογος επιτρέπει στο σύστημα να αυτοματοποιεί την ταυτοποίηση των τοποθεσιών σε κάθε κείμενο της συλλογής. Κατόπιν αυτά τα τοπωνύμια μεταφράζονται σε γεωγραφικές συντεταγμένες και αποτυπώνονται στον άτλαντα της PDL, που είναι ο μηχανισμός που επιτρέπει τις ερωτήσεις προς τα δεδομένα και την οπτικοποίηση των απαντήσεων (Chavez, 2000).

Το αρχικό σχήμα καλύπτει περίπου 5.000 αρχαίες τοποθεσίες. Οι πληροφορίες για το χώρο δηλαδή οι γεωγραφικές συντεταγμένες είναι αποθηκευμένες σε μια βάση δεδομένων PostGreSQL που επιτρέπει την εύκολη ενοποίηση με την εφαρμογή απεικόνισης MapServer* που χρησιμοποιείται. Η εφαρμογή MapServer παρέχει τα εργαλεία για την δημιουργία διαδραστικών διεπαφών μέσω του Διαδικτύου που αλληλεπιδρούν με την PDL και οπτικοποιούν τα χωρικά δεδομένα στο web site της PDL. (Chavez, 2000).

4.2 Περιγραφή του προβλήματος της γλωσσικής αμφισημίας και της αποσαφήνισης τοπωνυμίων.

Η αμφισημία είναι όρος της σημασιολογίας και χρησιμοποιείται για να δηλώσει λέξεις, φράσεις ή προτάσεις με διπλή, πολλαπλή ή διφορούμενη σημασία. Το πρόβλημα εξειδικεύεται στην αμφισημία κυρίων ονομάτων και τοπωνυμίων. Στις εφαρμογές επεξεργασίας κειμένου όπως συστήματα μηχανικής μετάφρασης, συστήματα ανάκτησης πληροφοριών ή συστήματα κατανόησης της φυσικής γλώσσας είναι απαραίτητη η ταυτοποίηση λέξεων και φράσεων που αναφέρονται σε κύρια ονόματα, οργανισμούς, τόπους και άλλες οντότητες. Επιπλέον όταν οι χρήστες μιας ψηφιακής βιβλιοθήκης αναζητούν συνδυασμένα πρόσωπα, αντικείμενα και αφηρημένες έννοιες η γλώσσα παρουσιάζει προβλήματα. Ακόμα και αν δεν δημιουργηθούν θησαυροί ή οντολογίες για τη συσχέτιση λέξεων (π.χ. truth (αγγλικά) – veritas (λατινικά) είναι αναγκαίο να αποσαφηνιστούν λεκτικές οντότητες όπως π.χ. το κύριο όνομα Αλέξανδρος (υπήρχαν 5 βασιλιάδες με αυτό το όνομα μόνο στη Μακεδονία) ή το τοπωνύμιο Springfield (πόλη του Illinois, του Missouri ή...των Simpsons).

Έτσι ένα τμήμα της βασικής αρχιτεκτονικής της PDL ήταν μια βάση δεδομένων οντοτήτων και των εναλλακτικών ονομάτων (που εμπλουτίστηκε με όρους θησαυρών και άλλων πληροφοριακών πηγών) με κωδικοποιημένες τις αντιστοιχίες των όρων έτσι ώστε οι συσχετίσεις των όρων να δίνουν ολοκληρωμένα αποτελέσματα π.χ. μια αναζήτηση με τον όρο Hercules να φέρει τεκμήρια και για τους όρους Heracles και Herakles. Τα ολοκληρωμένα αυτά αποτελέσματα περιλαμβάνουν όπως έχουμε αναφέρει και την οπτικοποιημένη παρουσίαση τους μέσω γεωγραφικής διεπαφής ώστε τα τοπωνύμια που περιέχονται στα αντικείμενα να «δημιουργούν» αυτόματα χάρτες. (Smith, 2000).

Ειδικότερα, τα τοπωνύμια που περιλαμβάνονται σε μια ψηφιακή βιβλιοθήκη παρουσιάζουν αρκετά προβλήματα στην επεξεργασία και αυτό γιατί:

* Το MapServer είναι ένα λογισμικό OpenSource για την κατασκευή εφαρμογών αποτύπωσης γεωγραφικών δεδομένων που επιτρέπουν στο χρήστη να οδηγείται μέσω του χάρτη στο περιεχόμενο (βλ. <http://mapserver.gis.umn.edu/>)

- τα ονόματα δεν είναι μοναδικά (π.χ. το όνομα San Jose είναι συνηθισμένο όνομα πόλης στην Κεντρική και Βόρεια Αμερική, υπάρχει η πόλη Alameda και η κομητεία Alameda και οι περιοχές Galicia (Πολωνία) και Galicia (Ισπανία)).
- οι τόποι και οι γεωγραφικές οντότητες αλλάζουν μέγεθος, σχήμα και όνομα μέσα στο χρόνο (π.χ. η Πρωσία δεν υπάρχει πια, η Γερμανία δεν υπήρχε ως κράτος πριν το 1870, η Σοβιετική Ένωση δεν υπάρχει πια, η Αυστρία είναι πολύ μικρότερη από την Αυστρο-Ουγγαρία, η Αγία Πετρούπολη ονομάστηκε Λένινγκραντ και πάλι Αγία Πετρούπολη, κτλ.)
- μια περιοχή μπορεί να έχει περισσότερα από ένα ονόματα σε διαφορετικές διαλέκτους και το ίδιο όνομα να έχει δύο ή περισσότερες μορφές έπειτα από μεταγραφή (π.χ. Peking, Beijing).
- Κάποια τοπωνύμια είναι απλά συμβάσεις κατασκευασμένες από συγγραφείς ή σχολιαστές σε μεμονωμένες περιπτώσεις και δεν έχουν σχέση με την επίσημη ονομασία του τόπου. (Larson, 1995).

Έτσι στην PDL τα τοπωνύμια που αναφέρονται στη Βόρεια Αμερική πολλές φορές είναι ονόματα προσώπων (π.χ. “John, Louisiana”). Όπως θα δούμε και παρακάτω για όσα ονόματα αναγνωριστούν ως τοπωνύμια το 90% αναφέρεται σε περισσότερα από ένα μέρη. (Smith, 2001).

4.3 Περιγραφή του προβλήματος και σχετικές εργασίες

Η σύνδεση λέξεων και φράσεων σε ένα τεκμήριο με τις θέσεις τους σε ένα χάρτη είναι μια διαδικασία δύο σταδίων:

- 1) ταυτοποίησης και κατηγοριοποίησης ονομάτων και
- 2) αποσαφήνισης όσων ονομάτων είναι τοπωνύμια με βάση ένα γεωγραφικό λεξικό.

Οι περισσότερες έρευνες στα θέματα αυτά ασχολήθηκαν με το 1ο στάδιο. Κύρια γιατί πολλά συστήματα «αναγνώρισης ονομασμένης οντότητας» (“named entity recognition”) (δηλαδή κατηγοριοποίησης ονομάτων) στοχεύουν γενικά στην ταξινόμηση οντοτήτων σε κατηγορίες (πρόσωπο, οργανισμός, ημερομηνία, τόπος, προϊόν, κτλ.) και λειτουργούν συνήθως χωρίς κάποιο κατάλογο γνωστών ονομάτων. Αντίθετα σε μια ψηφιακή βιβλιοθήκη γεωγραφικών δεδομένων είναι απαραίτητη η χρήση γεωγραφικών λεξικών γιατί απαιτείται και ένας κατάλογος τοπωνυμίων και μια πηγή πληροφοριών για τη θέση τους (τοποθεσία).

Υπάρχουν δύο γενικές στρατηγικές «αναγνώρισης ονομασμένης οντότητας» που έχουν παρουσιάσει δύο γνωστά συστήματα: το Nymble (BBN) και το Nominator (IBM) με επιτυχία 88% έως 90% στην αποσαφήνιση ονομάτων. Άλλα συστήματα όπως το GIPSY (Geo-Referenced Information Processing System) αντιστοιχούν όλα τα γεωγραφικά ονόματα ενός τεκμηρίου με όλες τις πιθανές συντεταγμένες χωρίς αποσαφήνιση με επιτυχία 75% ώστε να δώσει την εικόνα όλης της περιοχής που περιλαμβάνει όλους του τόπους που αναφέρονται στο κείμενο.

Ωστόσο, τα τεκμήρια πολλών ψηφιακών βιβλιοθηκών παρουσιάζουν προβλήματα κατά την αυτόματη ταυτοποίηση και αποσαφήνιση τοπωνυμίων. Οι περισσότερες εργασίες έχουν ασχοληθεί κύρια με ειδησεογραφικά κείμενα που περιέχουν χαρακτηριστικά που μειώνουν την γλωσσική αμφισημία (π.χ. το όνομα Bill Clinton χαρακτηρίζεται από τις λέξεις “President”, “Ms”, κτλ.), είναι σύντομα σε έκταση και εντοπισμένα σε θεματολογία. Επιπλέον, σε κάθε νέο κείμενο το όνομα π.χ. Bill Clinton θα αναφέρεται ολόκληρο (με τον τίτλο του) από την αρχή. Επίσης,

τα τοπωνύμια συνήθως πάντα έχουν μια ετικέτα που διαλύει την αμφισημία (π.χ. “Clinton, New Jersey”). Από την άλλη, οι ψηφιακές βιβλιοθήκες περιέχουν τεκμήρια μεγάλα σε έκταση, διαφορετικού ύφους, που αναφέρονται σε διάφορους τόπους και χρόνους (π.χ. ένα έργο για τον Shakespeare μπορεί να έχει αναφορές στο Stanford του 16^{ου} αι., στο Cambridge του 19^{ου} αι. και στο Berkeley του 20^{ου} αι.). Ή στην περίπτωση της Σάμου και της Μιλήτου: τώρα ανήκουν στην Ελλάδα και την Τουρκία αντίστοιχα αλλά και οι δύο ιδρύθηκαν από Ίωνες και απέχουν μόλις 30 km. Αυτό σημαίνει ότι το διευρυμένο ιστορικό πεδίο μιας συλλογής κάνει ένα γεωγραφικό λεξικό άχρηστο ή παραπλανητικό σε κάποιες περιπτώσεις.

Μια ψηφιακή βιβλιοθήκη με ετερογενές υλικό πρέπει να αντιμετωπίσει το πρόβλημα της σύγκρουσης ανάμεσα στους καταλόγους καθιερωμένων ονομάτων των διαφορετικών ειδών τεκμηρίων που περιλαμβάνει. Δηλαδή, οι αμφισημίες όρων προκαλούνται και από το πόσα και ποια γεωγραφικά λεξικά ή άλλες πηγές θα χρησιμοποιηθούν. Κάποιες τέτοιες ασάφειες φαίνονται εκ των προτέρων από τις κατανομές ενός γεωγραφικού λεξικού (Getty Thesaurus of Geographic Names) (Πίν.1). Έτσι σε περιοχές που κατοικούνται από παλιά (Ευρώπη, Ασία, Αφρική) ένα μέρος μπορεί να έχει πολλά ονόματα στη διάρκεια της ιστορίας αλλά συνήθως διαφορετικά. Αντίθετα, σε περιοχές που αποικίστηκαν σχετικά πρόσφατα όπως η Αμερική, πολλά μέρη έχουν το ίδιο όνομα. Παρόμοιες αμφισημίες προκύπτουν όταν άνθρωποι και τοποθεσίες έχουν το ίδιο όνομα. Αυτό είναι σχετικά σπάνιο για ελληνικά και λατινικά τοπωνύμια, πιο συχνό στη Βρετανία και απολύτως συνηθισμένο στην Αμερική.

Στην πράξη, το εύρος του προβλήματος της αποσαφήνισης τοπωνυμίων είναι πολύ μεγάλο: (χωρίς να υπολογίσουμε τα ονόματα που κατά λάθος μπορεί να θεωρηθούν ως τοπωνύμια) περίπου 92% των τοπωνυμίων της PDL αναφέρονται σε περισσότερα από ένα μέρη.

Πίν. 1

Ήπειρος	% τόπων με πολλά ονόματα	% ονομάτων που αντιστοιχούν σε πολλούς τόπους
B. & K. Αμερική	11,5	57,1
Ωκεανία	6,9	29,2
N. Αμερική	11,6	25,0
Ασία	32,7	20,3
Αφρική	2,0	18,2
Ευρώπη	18,2	16,6

5. Η διαδικασία αποσαφήνισης

Η αποσαφήνιση τοπωνυμίων συνίσταται στην ταυτοποίηση-κατηγοριοποίηση των ονομάτων και την αποσαφήνισή τους. Οι μέθοδοι της PDL για αυτές τις εργασίες βασίζονται σε εξωτερικά και εσωτερικά στοιχεία του κειμένου.

Τα εσωτερικά στοιχεία περιλαμβάνουν τιμητικές προσφωνήσεις και τίτλους, γενικούς γεωγραφικούς χαρακτηρισμούς ή το γλωσσολογικό περιβάλλον. Τα εξωτερικά στοιχεία περιλαμβάνουν γεωγραφικά λεξικά, βιογραφίες και γενικές γνώσεις γλωσσολογίας.

5.1 Η δημιουργία του γεωγραφικού λεξικού.

Προτού ξεκινήσει η ταυτοποίηση και η αποσαφήνιση συγκεντρώνονται οι πηγές που θα χρησιμοποιηθούν για να ληφθούν οι σχετικές αποφάσεις. Η PDL χρησιμοποιεί πηγές που έχουν κατασκευαστεί για το σκοπό της κωδικοποίησης γεωγραφικών δεδομένων όπως το Getty Thesaurus of Geographic Names, το Cruchley's gazetteer of London ή πηγές που περιλαμβάνονται στην PDL όπως το Dictionary of National Biography. Το γεωγραφικό λεξικό που κατασκευάστηκε με αυτό τον τρόπο περιλαμβάνει πάνω από 1εκ τοπωνύμια.

5.2 Κατηγοριοποίηση και ταυτοποίηση ονομάτων.

Κατόπιν γίνεται σάρωση των τεκμηρίων της ψηφιακής βιβλιοθήκης για πιθανά κύρια ονόματα τα οποία εντάσσονται σε μεγάλες κατηγορίες όπως: πρόσωπο, τόπος ή ημερομηνία. Αυτή η διαδικασία γίνεται με τη χρήση απλών ευρετικών μεθόδων και όχι με συστήματα μάθησης αφού δεν υπήρχαν επεξεργασμένα δείγματα δεδομένων (training data) για τους τύπους των τεκμηρίων της PDL αλλά και επειδή το ενδιαφέρον εστιάζεται στην ταυτοποίηση τοπωνυμίων και όχι στο ευρύτερο έργο της αναγνώρισης ονομασμένων οντοτήτων. Γενικά η διαδικασία της ανεύρεσης των κύριων ονομάτων είναι η ακόλουθη: Η δημιουργία ενός καταλόγου με όλα τα πιθανά κύρια ονόματα ενός κειμένου. Η εφαρμογή ευρετικών μεθόδων για τη διάσπαση των σύνθετων ονομάτων ή φράσεων. Η ομαδοποίηση όλων των τύπων (variants) κάθε ονόματος που αναφέρεται στην ίδια οντότητα. Τέλος η κατηγοριοποίηση των οντοτήτων με κάποιες προεπιλεγμένες κατηγορίες. Αυτή είναι η διαδικασία που ακολουθήθηκε από το σύστημα Nominator (Wacholder, 1998) ενώ παρόμοιο σύστημα κατηγοριοποίησης χρησιμοποίησε και η PDL. Ειδικότερα κάποια από τα στοιχεία που χρησιμοποιεί η PDL στα αγγλικά κείμενα είναι:

1) Ευρέως χρησιμοποιούμενα συστήματα στίξης και κεφαλαιογράμματος γραφής. Έτσι αρχικά ως υπονήφια κύρια ονόματα είναι: σειρές λέξεων που έχουν κεφαλαίο το αρχικό γράμμα και προτάσεις που οριοθετούνται από τελείες.

2) Διακριτικά που σημειώσει οι εκδότες των τεκμηρίων είτε αναφέροντας ένα όνομα ως όνομα προσώπου ή τύπου είτε άμεσα συνδέοντάς το με μια εγγραφή γεωγραφικού λεξικού.

3) Προσφωνήσεις ευγενείας και τίτλοι (π.χ. “Mrs”, “Dr”, “Col.”) αποτελούν ισχυρές ενδείξεις ότι το όνομα που ακολουθεί είναι όνομα προσώπου. Επιπλέον, αν , για παράδειγμα, εντοπιστεί μια αναφορά σε κάποιον “Col. Aldrich” κάθε επόμενη αναφορά στη λέξη Aldrich θα ταξινομηθεί αυτόματα ως όνομα προσώπου. Επίσης, γενικοί γεωγραφικοί προσδιορισμοί (π.χ. “Rocky Mountains”, “Charles River”) θεωρούνται ως ένδειξη ότι το όνομα είναι γεωγραφικό. Έτσι, μεμονωμένες εμφανίσεις των πιο συνηθισμένων ονομάτων που περιλαμβάνονται στο βιογραφικό λεξικό της PDL κατηγοριοποιούνται ως ονόματα προσώπων και όχι ως τοπωνύμια σε αυτή τη διαδικασία (αφού είναι μάλλον απίθανο απλώς το όνομα “John” να αναφέρεται σε κάποια πόλη) .

5.2.1 Αποσαφήνιση πιθανών τοπωνυμίων.

Αφού το σύστημα έχει κατηγοριοποιήσει έτσι τα κύρια ονόματα σε κατηγορίες προσπαθεί να συνταιριάζει τα ονόματα που ταξινομήθηκαν ως

γεωγραφικά (αλλά και αυτά για τα οποία υπάρχει αμφιβολία) με τους όρους ενός γεωγραφικού λεξικού ώστε τελικά να συνδεθούν με γεωγραφικές συντεταγμένες (αφού θα είχε μικρή χρησιμότητα να προσδιοριστούν όροι που δεν θα συνδέονταν με τέτοιες συντεταγμένες). Όπως αναφέρθηκε ήδη για τα ονόματα που περιλαμβάνονται στη συλλογή και έχουν τουλάχιστο μια αντιστοιχία με εγγραφή του γεωγραφικού λεξικού, το 92% αναφέρεται σε περισσότερες από μια οντότητες.

Η διαδικασία της αποσαφήνισης των πιθανών τοπωνυμίων που αντιστοιχούν σε μια θέση προχωρά βασισμένη στα παρακάτω τρία στοιχεία :

- β) Στα συμφραζόμενα του τεκμηρίου (document context) ή γενικό χαρακτηρισμό του τεκμηρίου.
- α) Στα εντοπισμένα συμφραζόμενα (local context).
- γ) Στη γενική γνώση του κόσμου.

Ειδικότερα:

- α) Τα συμφραζόμενα του τεκμηρίου αναφέρονται στην αριθμητική επικράτηση γεωγραφικών αναφορών σε ολόκληρο το τεκμήριο και δίνουν ένα γενικό χαρακτηρισμό για όλο το τεκμήριο. Για σύντομα κείμενα τα εντοπισμένα συμφραζόμενα και αυτά ολόκληρου του τεκμηρίου θεωρούνται ταυτόσημα.
- β) Η πιο απλή περίπτωση εντοπισμένων συμφραζόμενων είναι οι ρητά αποσαφηνισμένες ετικέτες ή προσδιορισμοί (tags) που έχουν σημειώσει οι ίδιοι οι συγγραφείς δίπλα σε κάθε όνομα (π.χ. “Vienna, Austria”). Γενικά, ένας τύπος είναι πιο πιθανό να βρίσκεται κοντά σε άλλους τύπους που αναφέρονται στο κείμενο (π.χ. το “Lancaster” είναι πιο πιθανό να είναι αυτό της Pennsylvania και όχι αυτό της Arizona αν στο κείμενο υπάρχουν αναφορές και στη «Φιλαδέλφεια» και στο “Harrisburg” [σημ.: πόλεις της Pennsylvania]). Δηλαδή στόχος είναι να υπολογιστεί η χωρική εγγύτητα των τοποθεσιών που αναφέρονται στην ίδια πρόταση ή παράγραφο.
- γ) Η γενική γνώση του κόσμου αναφέρεται στις πληροφορίες που μπορούμε να πάρουμε από γεωγραφικά λεξικά ή άλλα πληροφοριακά έργα και συγκροτείται από στοιχεία όπως: γεωγραφικές συντεταγμένες ενός τόπου, διοικητική ή πολιτική οντότητα στην οποία ανήκει, το μέγεθός του και τη σπουδαιότητά του.

Το σύστημα ξεκινά παράγοντας ένα απλό χαρακτηρισμό για το γενικό περιεχόμενο του τεκμηρίου. Αυτό γίνεται με τον ακόλουθο τρόπο: Όλες οι πιθανές τοποθεσίες για όλα τα τοπωνύμια ομαδοποιούνται σε ένα πίνακα συντεταγμένων 1:1 (one-by-one degree grid) με βάρη που αποδίδονται στον αριθμό των αναφορών κάθε τοπωνυμίου. Το σύστημα περιορίζει την έκταση κάποιων πιθανοτήτων βασισμένο σε στοιχεία γενικής γνώσης. Κατόπιν, υπολογίζεται το κέντρο βάρους (centroid)* αυτού του πίνακα βαρών (weighted map) καθώς και η τυπική απόκλιση (standard deviation)** της απόστασης των διαφόρων σημείων από το κέντρο βάρους (centroid). Έπειτα αφαιρούνται τα σημεία που απέχουν περισσότερο από δύο τυπικές αποκλίσεις από το κέντρο βάρους και υπολογίζεται ένα νέο κέντρο βάρους του κειμένου από τα εναπομείναντα σημεία, αν υπάρχουν.

* centroid = κέντρο βάρους τριγώνου ή το σημείο τομής των τριών διαμέσων ενός τριγώνου με πλευρές Α.Β.Γ. Επιπλέον κέντρο βάρους ενός cluster.

** standard deviation = τυπική απόκλιση = η τετραγωνική ρίζα της διασποράς μιας τυχαίας μεταβλητής

5.3 Τελική αποσαφήνιση.

Στην επόμενη φάση γίνεται η τελική αποσαφήνιση των πιθανών τοπωνυμίων. Βρίσκονται τα εντοπισμένα συμφραζόμενα (local context) της εμφάνισης ενός τοπωνυμίου τα οποία παριστάνονται με ένα «κινούμενο παράθυρο» (moving window) που αποτελείται από τα τέσσερα τοπωνύμια που αναφέρονται πριν από αυτό και τα τέσσερα που αναφέρονται μετά. Αυτά τα εντοπισμένα συμφραζόμενα αποτελούνται μόνο από σαφή ή ήδη αποσαφηνισμένα τοπωνύμια. Κάθε πιθανή τοποθεσία για ένα τοπωνύμιο παίρνει μια βαθμολογία που βασίζεται: α) στην εγγύτητα της θέσης με άλλα τοπωνύμια γύρω της, β) στην εγγύτητά της με το κέντρο βάρους του τεκμηρίου, και γ) στη σχετική της σπουδαιότητα (π.χ. αν όλα τα υπόλοιπα στοιχεία είναι ίδια, μια χώρα βαθμολογείται υψηλότερα από μια πόλη). Σε αυτό το στάδιο, το σύστημα απορρίπτει ως πιθανά λάθος τις τοποθεσίες που δεν έχουν κάποιο σαφές χαρακτηριστικό αποσαφήνισης, που λαμβάνουν χαμηλή βαθμολογία σπουδαιότητας και που απέχουν πολύ από το κέντρο βάρους του τεκμηρίου. Αν δεν εξαλειφθεί με αυτό τον τρόπο, η πιθανή τοποθεσία ενός τοπωνυμίου με την υψηλότερη βαθμολογία θεωρείται νικήτρια. Αφού τελειώσει η εργασία του συστήματος αποσαφήνισης, τα τοπωνύμια «φορτώνονται» σε μια σχεσιακή βάση δεδομένων στην οποία έχει πρόσβαση η ψηφιακή βιβλιοθήκη.

6. Αξιολόγηση του συστήματος.

Η απόδοση του συστήματος αξιολογήθηκε με μεθόδους μέτρησης της ακρίβειας και ανάκτησης (precision and recall methods). Ποιοτικά, το σύστημα αποδίδει ικανοποιητικά στην παραγωγή χαρτών από τα μέρη που αναφέρονται σε ένα ολόκληρο τεκμήριο ή σε μια συλλογή καθώς και στην ανεύρεση αναφορών ενός συγκεκριμένου μέρους.

Για μια πιο λεπτομερειακή εξέταση του συστήματος σε διάφορα κείμενα, ένας ερευνητής εργάστηκε με το 20% των αποτελεσμάτων ενός κειμένου από κάθε μια από τις πέντε αντιπροσωπευτικές συλλογές της PDL (αρχαία Ελλάδα, αρχαία Ρώμη, συλλογή Bolles για το Λονδίνο, και δύο συλλογές της Βιβλιοθήκης του Κονγκρέσου για τον αποικισμό της Καλιφόρνιας και των άνω Μεσοδυτικών Πολιτειών). Με τη χρήση ενός μεγάλου γεωγραφικού λεξικού, το σύστημα τείνει περισσότερο προς την ανάκτηση. Στον Πιν. 2 φαίνονται τα αποτελέσματα για την ακρίβεια του συστήματος συνολικά (ότι «βλέπει» τελικά ο χρήστης), και για το σύστημα αποσαφήνισης τοπωνυμίων ειδικά ανεξαρτήτως της κατηγοριοποίησης ονομάτων. Επίσης φαίνεται το μέτρο F για όλο το σύστημα *.

Από στα στοιχεία φαίνεται ότι παρόλο που ο απλός αλγόριθμος ευρετικής κατηγοριοποίησης ήταν λιγότερο κατάλληλος για συγκεκριμένες περιπτώσεις, το ίδιο το σύστημα αποσαφήνισης τοπωνυμίων είχε καλά αποτελέσματα. Η κατηγοριοποίηση είχε καλύτερα αποτελέσματα στα κείμενα ελληνικής και ρωμαϊκής

* Το μέτρο F του συστήματος είναι ο βαθμός που συνδυάζει την ανάκτηση R και την ακρίβεια P με ένα συντελεστή βάρους ανάκτησης/ακρίβειας B^2 που είναι συνήθως 1:

$$F = \frac{(\beta^2 + 1)RP}{(\beta^2 R) + P}$$

Πιν. 2

Συλλογή	Ακρίβεια	Σύστημα αποσαφήνισης	Ανάκληση	F-measure
Αρχαία Ελλάδα	0,93	0,98	0,99	0,96
Αρχαία Ρώμη	0,91	0,99	1,00	0,95
Λονδίνο	0,86	0,92	0,96	0,91
Καλιφόρνια	0,83	0,92	0,96	0,89
Άνω Μεσοδ. Πολιτείες	0,74	0,89	0,89	0,81

ιστορίας παρά σε αυτά των άλλων περιοχών. Αυτό αντανακλά και το βαθμό αμφισημίας των τοπωνυμίων σε σχέση με άλλα ονόματα (βλ. Πίν. 1). Η αξιολόγηση έφερε στην επιφάνεια άλλο ένα γλωσσολογικό ζήτημα: όλες οι λαθεμένες ταυτοποιήσεις τοπωνυμίων στο ρωμαϊκό κείμενο (Ceasar's Gallic War*) σχετίζονταν με τους Γερμανούς (Germans) με τους οποίους πολεμούσε ο Καίσαρας. Η λέξη "German" υπάρχει στο γεωγραφικό λεξικό στην εγγραφή για τη Γερμανία αλλά η λέξη στον πληθυντικό δεν υπάρχει. Αυτό το ελάττωμα θα μπορούσε να διορθωθεί με αποκοπή των καταλήξεων από την αρχή αλλά επειδή γενικά τα κύρια ονόματα δεν κλίνονται στα αγγλικά κάτι τέτοιο θα προκαλούσε περισσότερο κακό παρά καλό. Αυτές οι μορφές των τοπωνυμίων μπορούν να προστεθούν στο γεωγραφικό λεξικό με το χέρι. Γενικά, ένα μεγάλο γεωγραφικό λεξικό με πάνω από 1 εκ. ονόματα συμπίπτει την ακρίβεια περισσότερο από κάθε άλλο παράγοντα.

7. Περιορισμοί του συστήματος και μελλοντικοί στόχοι.

Αν και το σύστημα αποσαφήνισης λειτουργεί ικανοποιητικά ο στόχος είναι η βελτίωση του συστήματος κατηγοριοποίησης κύρια για κείμενα της Β. Αμερικής. Πολλές προσεγγίσεις στην κατηγοριοποίηση απαιτούν επεξεργασμένα δείγματα δεδομένων (training data) ώστε το σύστημα να μάθει κανόνες συμφραζόμενων για την εμφάνιση των διαφόρων τύπων επώνυμων οντοτήτων.

Ένα άλλο σημείο βελτίωσης της απόδοσης του συστήματος θα είναι ο περιορισμός των διαθέσιμων τοπωνυμίων ενός κειμένου με βάση τη χρονική περίοδο. Ωστόσο, σε μια ανομοιογενή ψηφιακή βιβλιοθήκη ιστορικού περιεχομένου μπορεί να εμφανίζονται πολλές χρονικές αναφορές πολύ κοντά ή μια στην άλλη. Θα ήταν δυνατό να χρησιμοποιηθεί η επικράτηση τέτοιων αναφορών όπως γίνεται τώρα με τη χρήση του πίνακα βαρών για τις αναφορές χώρων ώστε να ιεραρχούνται οι πιθανότητες για κάθε τοπωνύμιο. Για παράδειγμα, η λέξη "Ovid" (Οβίδιος) σε ένα κείμενο ρωμαϊκής ποίησης είναι απίθανο να αναφέρεται στην πόλη Ovid του Idaho. Ενώ το σύστημα ως έχει τώρα το συμπεραίνει αυτό από την απόσταση της πόλης από την Ιταλία (όπου βρίσκονται τα περισσότερα μέρη που αναφέρονται στο κείμενο) το γεγονός ότι η πόλη των Η.Π.Α. ιδρύθηκε το 19^ο αι. θα μπορούσε να την αποκλείσει για ένα τεκμήριο που οι περισσότερες ημερομηνίες που περιλαμβάνει είναι του 1^{ου} αι. π.Χ. και του 1^{ου} αι. μ.Χ.

* («Απομνημονεύματα του Γαλατικού Πολέμου»)

Άλλο σημείο προβληματισμού είναι η αποκλειστική χρήση του «κέντρου βάρους» (centroid) που αναφέρεται στις περιοχές με τις περισσότερες αναφορές σε ένα τεκμήριο και που χαρακτηρίζει το περιεχόμενό του. Ο πειραματισμός που ήδη γίνεται αφορά τη χρήση στο χάρτη ενός ορθογωνίου (bounding rectangle) ή πολυγώνου που θα απεικονίζει την περιοχή ενδιαφέροντος για ένα τεκμήριο.

Τέλος, μια εργασία που ήδη έχει αρχίσει είναι η συγκρότηση ενός γεωγραφικού λεξικού ελληνικών και λατινικών τοπωνυμίων για κείμενα τα οποία δεν είναι στην αγγλική γλώσσα. Μεγάλο μέρος του θα μπορούσε να προέλθει από ψηφιοποιημένα λεξικά όπως το Harper's Dictionary of Classical Antiquities και το Smith's Dictionary of Greek and Roman Geography συν λέξεις με αποκοπή των καταλήξεων για αυτές τις λέξεις που κλίνονται.

8. Σχετικές εργασίες.

8.1 Alexandria Digital Library (ADL)

Η ADL είναι μια διαμοιρασμένη ψηφιακή βιβλιοθήκη με δεδομένα γεωγραφικού χαρακτήρα δηλαδή με τεκμήρια (holdings) που συνδέονται με μια ή περισσότερες περιοχές (footprints) στην επιφάνεια της γης. Ένας από τους στόχους είναι οι χρήστες να απευθύνουν μόνο ένα ερώτημα προς όλη την ετερογενή συλλογή πράγμα που σημαίνει ότι υποστηρίζει πολλά αναζητήσιμα σχήματα μεταδεδομένων. Αυτό το κατόρθωσε με την παραγωγή ενός μικρού set χαρακτηριστικών μεταδεδομένων που αποτελείται από 8 στοιχεία και καλείται "search bucket". Όσων αφορά τα footprints (δηλαδή τα «αποτυπώματα» των γεωγραφικών δεδομένων), αυτά μπορούν να συνδεθούν με κάθε είδους πληροφορία για μια τοποθεσία όπως χάρτη, αεροφωτογραφία, μουσειακά αντικείμενα, μουσική, κείμενο. Έτσι η ADL είναι μια «γεω-βιβλιοθήκη» που το κύριο χαρακτηριστικό των αντικειμένων της είναι η θέση τους στη Γη δηλαδή το «αποτύπωμά» τους.

Στην ADL έχει δοθεί ιδιαίτερη σημασία στο ρόλο του ψηφιακού γεωγραφικού λεξικού και αυτό γιατί: 1) το γεωγραφικό λεξικό απαντά στην ερώτηση «που είναι κάτι», 2) συνδέει γεωγραφικά ονόματα με τις θέσεις τους και επιτρέπει στο χρήστη να βρει αντικείμενα μέσα από το ταίριασμα του αποτυπώματος ενός γεωγραφικού ονόματος με το αποτύπωμα ενός αντικειμένου, 3) επιτρέπει στο χρήστη να εντοπίσει τύπους γεωγραφικών χαρακτηριστικών σε μια ομοιογενή περιοχή. Το γεωγραφικό λεξικό της ADL δημιουργήθηκε από τη συνένωση των διαθέσιμων πηγών GNIS (Geographic Names Information System) και GNPS (Geographic Names Processing System). Τα προβλήματα που προέκυψαν είχαν να κάνουν κυρίως με τη χρήση διαφορετικών κατηγοριών γεωγραφικών χαρακτηριστικών αυτών των λεξικών αλλά και με το ότι τα γεωγραφικά ονόματα είτε περιέχουν τα ίδια πληροφορίες για τον τύπο τους (π.χ. Perth Airport) είτε δεν περιέχουν αρκετή τέτοια πληροφορία (π.χ. "California" και όχι "State of California"). Επίσης άλλοι προσδιοριστικοί όροι είναι κοινή για διαφόρους τύπους τοποθεσιών πράγμα που αυξάνει την αμφισημία (π.χ. το "Kindley Field" αναφέρεται σε αεροδρόμιο και το "Bilate River Field" αναφέρεται σε ηφαίστειο).

Τελικά η δομή της βάσης δεδομένων του γεωγραφικού λεξικού που κατασκευάστηκε είναι απλή και η κάθε εγγραφή αποτελείται από χαρακτηριστικά (όνομα, κατηγορία, τύπος, συντεταγμένες, κτλ.) που επιτρέπουν στο χρήστη να αναζητήσει γεωγραφικά ονόματα ή τύπους γεωγραφικών χαρακτηριστικών με αποτέλεσμα τις θέσεις τους πάνω στο χάρτη.

Για περισσότερες πληροφορίες στα (Hill, et al. 1999) και (Hill, 2000).

8.2 GIPSY (Geo-Referenced Information Processing System)

Πρόκειται για ένα μοντέλο ευρετηρίασης γεωγραφικών όρων που βρίσκονται σε κείμενα. Η στρατηγική του συστήματος είναι η εξόρυξη τοπωνυμίων και γενικών γεωγραφικών χαρακτηριστικών από τεκμήρια και η παράλληλη χρήση τους για τον υπολογισμό της περιοχής στην οποία αναφέρεται το τεκμήριο.

Το GIPSY χρησιμοποιεί ένα αλγόριθμο τριών σταδίων που βασίζεται σε ένα θησαυρό ή γεωγραφικό λεξικό.

1^ο στάδιο: Ταυτοποίηση γεωγραφικών τοπωνυμίων και φράσεων.

2^ο στάδιο: Εντοπισμός συναφών δεδομένων.

3^ο στάδιο: Χρήση επικαλυπτόμενων πολυγώνων για τον υπολογισμό των πιθανών τοποθεσιών.

Αναλυτικά το σύστημα περιγράφεται στα (Larson, 1996) και (Woodruff and Plaunt, 1994).

8.3 Nominator (IBM T.J. Watson Research Center)

Σύστημα αναγνώρισης κυρίων ονομάτων που ασχολήθηκε ιδιαίτερα με τους τύπους γλωσσικής αμφισημίας – δομική και εννοιολογική – που καθιστούν δύσκολη των εξεύρεση κυρίων ονομάτων (προσώπων, τόπων, οργανισμών, κτλ.) μέσα στο κείμενο. Επίσης ασχολήθηκε και με τις ευρετικές μεθόδους που χρησιμοποιήθηκαν για την αποσαφήνιση αυτών των ονομάτων. Γενικά εφαρμόζει ένα σύνολο ευρετικών μεθόδων σε καταλόγους λέξεων ή φράσεων στη βάση τύπων στίξης, κεφαλαιογράμματος γραφής και θέσης μέσα σε μια πρόταση ή τεκμήριο προκειμένου αρχικά να ανακαλύψει τα κύρια ονόματα και κατόπιν να λύσει τις δομικές και σημασιολογικές αμφισημίες. Αυτό το πετυχαίνει εξετάζοντας τη συνύπαρξη κυρίων ονομάτων στο ίδιο κείμενο και κύρια τις επαναλήψεις διαφόρων μορφών της ίδια λέξης (π.χ. αρκτικόλεξα).

Η αξιολόγηση του συστήματος έγινε σε 88 τεκμήρια της εφημερίδας Wall Street Journal και έδωσε ποσοστό επιτυχίας 92% στην ακρίβεια ταυτοποίησης και 79% στην ακρίβεια αποσαφήνισης. Αναλυτικά το σύστημα περιγράφεται στο (Wacholder, 1998).

8.4 Nymble

Πρόκειται για σύστημα αναγνώρισης ονομασμένων οντοτήτων (named entity recognition system) με τη χρήση μιας εκδοχής του μοντέλου πιθανοτήτων HMM (Hidden Markov Model). Τύποι αυτού του μοντέλου έχουν χρησιμοποιηθεί από διάφορες ερευνητικές κοινότητες που εργάζονται στον τομέα της επεξεργασίας της φυσικής γλώσσας. Πρόκειται βασικά για μια στατιστική προσέγγιση μάθησης για την εξεύρεση ονομάτων και άλλων μη-επαναλαμβανόμενων οντοτήτων μέσα σε κείμενο με τη χρήση του συγκεκριμένου στατιστικού μοντέλου. Η αξιολόγηση του Nymble έδειξε ότι ξεπερνά σε απόδοση όλα τα αντίστοιχα συστήματα μάθησης για εξεύρεση ονομάτων. Επιπλέον, έχει πάνω από 90% ποσοστό ακρίβειας έτσι ώστε να θεωρείται ότι πετυχαίνει “near-human performance”.

Για περισσότερες πληροφορίες (Bikel, et al., 1998).

9. Η χρησιμότητα του συστήματος και πιθανές εφαρμογές στον ελληνικό χώρο. Ερωτήματα και προβληματισμοί.

Εξ αιτίας της φύσης και των χαρακτηριστικών τους οι ψηφιακές βιβλιοθήκες αποτελούν φορείς συγκέντρωσης επεξεργασίας και παρουσίασης ψηφιακών δεδομένων που απευθύνονται σε ερευνητές με πολύ εξειδικευμένα αντικείμενα ή ερευνητές που χρειάζονται ειδικού τύπου πληροφορίες. Τέτοιες πληροφορίες είναι και αυτές που σχετίζονται με γεωγραφικά δεδομένα ή που τα δεδομένα τους περιέχουν και χωρική διάσταση. Οι ερευνητές των φυσικών επιστημών ασχολούνται με τέτοιους είδους δεδομένα αλλά όλο και μεγαλώνει το ενδιαφέρον και από ερευνητές των κοινωνικών επιστημών για την ανάδειξη των μεταβλητών του χώρου και την επίδρασή τους στα φαινόμενα που μελετούν. Έτσι, οι ιστορικοί, κοινωνιολόγοι, εθνογράφοι, ερευνητές της περιφερειακής ανάπτυξης και άλλοι μελετούν τις κατανομές στο χώρο των αντικειμένων τους. Είναι φανερό πως υπάρχει ένα τεράστιο ερευνητικό πεδίο ανεξερεύνητο ακόμα από διάφορους κλάδους και ειδικά στον ελληνικό χώρο που τώρα ξεκινάμε να εκμεταλλευτούμε τις νέες δυνατότητες διαδικτύου.

Το πρόβλημα της αποσαφήνισης τοπωνυμίων που παρουσιάσαμε είναι μόνο μια πλευρά ενός από τα δεκάδες προβλήματα που θα αντιμετωπίσει μια προσπάθεια δημιουργίας μιας ψηφιακής βιβλιοθήκης που ειδικεύεται στις ανθρωπιστικές επιστήμες και έχει και γεωγραφική παράμετρο. Το συγκεκριμένο πρόβλημα που παρουσιάσαμε θα αντιμετωπιστεί φυσικά και από οποιαδήποτε προσπάθεια ξεκινήσει και στην Ελλάδα καθώς θα προκύψει η ανάγκη αποσαφήνισης των διαφορετικά τοπωνυμίων για το ίδιο μέρος (διαφορετικές γραφές, διαφορετικά ονόματα), με τα ίδια τοπωνύμια που αναφέρονται σε διαφορετικά μέρη, με τις αλλαγές ονομάτων με το πέρασμα του χρόνου, με τα διαφορετικές γλωσσικής προέλευσης ονόματα (τουρκικά, σλαβικά, κτλ.) που συναντάμε στον ελληνικό χώρο, κ.ά. Αυτά σημαίνουν ανάπτυξη συστημάτων κατηγοριοποίησης και αποσαφήνισης ονομάτων αντίστοιχων με αυτά του εξωτερικού. Απαραίτητη επιπλέον είναι και η σύνταξη ενός ψηφιακού γεωγραφικού λεξικού που λείπει αυτή τη στιγμή (προς το παρόν υπάρχει μόνο έντυπη έκδοση του 1981 από τη Γεωγραφική Υπηρεσία Στρατού). Το έργο της ψηφιακής βιβλιοθήκης Perseus (αλλά και άλλων αντίστοιχων βιβλιοθηκών όπως της Alexandria Digital Library) μπορεί να αποτελέσει ένα υπόδειγμα για αντίστοιχες προσπάθειες στον ελληνικό χώρο αφού έχει ως κύριο αντικείμενο ένα θέμα και μια ιστορική-χρονική περίοδο που μας αφορά άμεσα.

Θα πρέπει να σημειώσουμε ότι έως σήμερα δεν υπάρχουν πολλά παραδείγματα δημιουργίας ψηφιακών βιβλιοθηκών ιστορικού περιεχομένου εκτός ίσως από το πρόγραμμα Βάση Γνώσεων για τον Ελληνισμό της Μικράς Ασίας του Ιδρύματος Μείζονος Ελληνισμού που περιλαμβάνει τη δημιουργία Αρχείου Τοπωνυμίων το οποίο έχει ως στόχους: 1) τη δημιουργία ενός εξαντλητικού θησαυρού ονομάτων, 2) την τοποθέτηση των υπό εξέταση αντικειμένων στο χώρο, 3) τη δημιουργία μιας βάσης δεδομένων με ιστορικές και ανθρωπογεωγραφικές πληροφορίες για την ελληνική Μικρά Ασία στη διαχρονία.

Αυτή και παρόμοιες εφαρμογές θα απαιτήσουν ανάμεσα σε άλλα και την αντιμετώπιση του προβλήματος της γλωσσικής ανάλυσης και της αποσαφήνισης των γεωγραφικών όρων ώστε να εκμεταλλευτούν τις δυνατότητες που δίνονται από τα γεωγραφικά συστήματα πληροφοριών και τα απεικονιστικά συστήματα για την κατασκευή χαρτών. Το ζητούμενο αποτέλεσμα είναι η πληρέστερη και εναργέστερη

παρουσίαση των αντικειμένων αλλά και επιπλέον ο ανασυνδυασμός αυτών των αντικειμένων για την δημιουργία νέας γνώσης.

Υπάρχουν φυσικά δεκάδες ερευνητικά πεδία που θα μπορούσαν να ωφεληθούν από παρόμοιες ή αντίστοιχες εφαρμογές. Παραδείγματος χάριν:

- Ψηφιακή βιβλιοθήκη για τη μελέτη του ελληνικού εμφυλίου πολέμου με απεικονίσεις των πεδίων μαχών, των περιοχών ενδιαφέροντος των αντιμαχόμενων κτλ.

- Ψηφιακή βιβλιοθήκη για την παρουσίαση του θέματος της μετανάστευσης στο Βαλκανικό χώρο μετά την πτώση του Τείχους και την κατάρρευση των ανατολικών καθεστώτων.

- Ψηφιακή βιβλιοθήκη για τη μελέτη της εγκληματικότητας στο αστικό περιβάλλον της Αθήνας, κ.ο.κ.

Συμπερασματικά θα μπορούσαμε να πούμε ότι μια σχετική προσπάθεια θέτει καινούργιες προκλήσεις όχι μόνο για τους επιστήμονες της πληροφόρησης και τους ειδικούς των πληροφοριακών συστημάτων αλλά κύρια για τους βιβλιοθηκονόμους που θα πρέπει :

- Να συνεργαστούν με επιστήμονες διαφόρων επιστημονικών τομέων για τον καθορισμό του εύρους μιας ανάλογης προσπάθειας και τους σκοπούς της.

- Να ερευνήσουν τα διαθέσιμα εργαλεία επεξεργασίας τεκμηρίων (και ειδικά της γλώσσας), αναζήτησης, ανάκτησης και παρουσίασης των δεδομένων ή να αναπτύξουν νέα.

- Να εξασφαλίσουν τη συνέχεια ενός έργου που έχει χαρακτήρα ιστορικό και προαγωγής της εθνικής κληρονομιάς.

10. ΒΙΒΛΙΟΓΡΑΦΙΑ

- Beard, K., Smith, T. & Hill, L. (1997). Meta-information models for georeferenced digital library collections. Proceedings of the Second IEEE Metadata Conference, September 16-17, 1997. Available at: <http://www.computer.org/proceedings/meta97/papers/kbeard/kbeard.html> (last accessed at 23 January 2004).
- Bikel, D.M., Miller, S., Schwartz, R., & Weischedel, R. (1998). Nymble: a high performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pp. 194-201. Available at: <http://acl.ld.upenn.edu/A/A97/-1029.pdf> (last accessed at 23 January 2004).
- Chavez, Robert F. (2000). Generating and reintegrating geospatial data. Proceeding of the Fifth ACM Conference on Digital Libraries (San Antonio, June 2-7, 2000). New York: ACM Press. Available at: <http://portal.acm.org/citation.cfm?id=336684> (last accessed at 23 January 2004).
- Crane, G. (1998, January). The Perseus Project and beyond: how building a digital library challenges the humanities and technology. D-Lib Magazine. Available at: <http://www.dlib.org/dlib/january98/01crane.html> (last accessed at 11 December 2003).
- Crane, G. (1996). Building a digital library: the Perseus Project as a case study in the humanities. In Proceedings of the 1st ACM International Conference on Digital Libraries, pp. 3-10, 1996.
- Frew, J., Freeston, M., Hill, L., Janec, G., Larsgaard, M., Zheng, Q. (1999). Generic query metadata for geospatial digital libraries. Available at: <http://www.computer.org/proceedings/meta/1999/papers/55/jfrew.htm> (last accessed at 23 January 2004).
- Getty Thesaurus of Geographic Names on line (2000). Available at: http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (last accessed at 29 January 2004).
- Hill, Linda L. (2000). Core elements of Digital Gazetteers: placenames, categories and footprints. Available at: http://www.alexandria.ucsb.edu/~hill/paper_drafts/ECDL2000_paperdraft17.pdf (last accessed at 23 January 2004).
- Hill, L.L., Frew, J., & Zheng, Q. (1999, January). Geographic names: the implementation of a gazetteer in a georeferenced digital library. D-Lib Magazine. Available at: <http://www.dlib.org/january99/hill/o1hill.html> (last accessed at 9 December 2003).
- Larson, Ray R. (1996). Geographic information retrieval and spatial browsing. In GIS and Libraries: Patrons, maps and spatial information, Linda Smith and Myke Gluck, Eds., University of Illinois, pp. 81-124. Available at: http://sherlock.berkeley.edu/geo_ir/PART1.html (last accessed at 23 January 2004).

Mahoney, Anne (2000). Explicit and implicit searching in the Perseus Digital Library. Eleventh ACM Conference on Hypertext and, June 2000. Available at: http://www.ics.mq.edu.au/~einat/info_doors/program/Mahoney.html (last accessed at 11 December 2003).

Mahoney, A., Rydberg-Cox, J., Smith, D.A., Wulfman, C.E. (2000). Generalizing the Perseus XML Document Manager. In *Linguistic Exploration: Workshop on Web-based Language Documentation and Description*, 12-15 December 2000, Philadelphia. Available at: <http://www ldc.upenn.edu/exploration/exlp2000/papers/mahoney/mahoney.htm> (last accessed at 30 January 2004)

MapServer Homepage. Available at: <http://mapserver.gis.umn.edu/> (last accessed at 5 January 2004).

Rydberg-Cox, J.A., Chavez, R.F., Smith, D.A., Mahoney, A., & Crane, G.R. (2000). Knowledge management in the Perseus Digital Library. *Ariadne*, 25. Available at: <http://www.ariadne.ac.uk/issue25/rydberg-cox/> (last accessed at 11 December 2003).

Smith, David A. (2002). Detecting events with date and place information in unstructured text. In Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries, pp. 191-196, Portland, OR, July 2002. Available at: <http://www.perseus.tufts.edu/Articles/datestat.pdf> (last accessed at 23 January 2004).

Smith, David A. (2001). Linking and gathering: automated hypertext in the Perseus Digital Library. Proceedings of 2001 Joint International Conference of the Association for Computers and the Humanities, June 13-16, 2001. Available at: http://www.nyu.edu/humanities/ach_allc2001/papers/smith-david/ (last accessed at 11 December 2003).

Smith, D.A., Crane, G. (2001). Disambiguating geographic names in a historical digital library. In P. Constantopoulos and I.T. Solvberg (eds.): ECDL 2001, LNCS 2163, (pp.127-136), Berlin: Springer-Verlag, 2001. Available at: <http://www.perseus.tufts.edu/Articles/geodl01.pdf> (last accessed at 23 January 2004).

Smith, D.A., Mahoney, A., Rydberg-Cox, J.A. (2000b). Management of XML Documents in an integrated digital library. *Proceedings of Extreme Markup Languages 2000*, 219-224. Available at: <http://www.perseus.tufts.edu/Articles/hopper.pdf> (last accessed at 30 January 2004).

Smith, D.A., Rydberg-Cox, J.A., & Crane, G.R. (2000a). The Perseus Project: a digital library for the humanities. *Literary and Linguistic Computing*, 15(1), 15-25

USGS. Geographic Names Information System (GNIS). (2003). Available at: <http://geonames.usgs.gov> (last accessed at 29 January 2004).

Wacholder, N., Ravin, Y., & Choi, M. (1997). Disambiguation of proper names in text. In *Proceeding of the Fifth Conference on Applied Natural Language Processing* [1], pp. 202-208. Available at: <http://www.cs.mu.az.au/acl/A/A97/A97-1030.pdf> (last accessed at 23 January 2004).

Woodruff, A.G., Plaunt, C. (1994). GIPSY: Georeferenced Information Processing System. *Journal of the American Society for Information Science*, 45(9), 645-655.

Ίδρυμα Μείζονος Ελληνισμού. Αρχείο Τοπωνυμίων. Διαθέσιμο στη διεύθυνση: <http://www.ime.gr/fhw.gr/projects/databank/geodb.html> (ημερομηνία πρόσβασης 21 Ιανουαρίου 2004).

11. ΠΑΡΑΡΤΗΜΑΤΑ

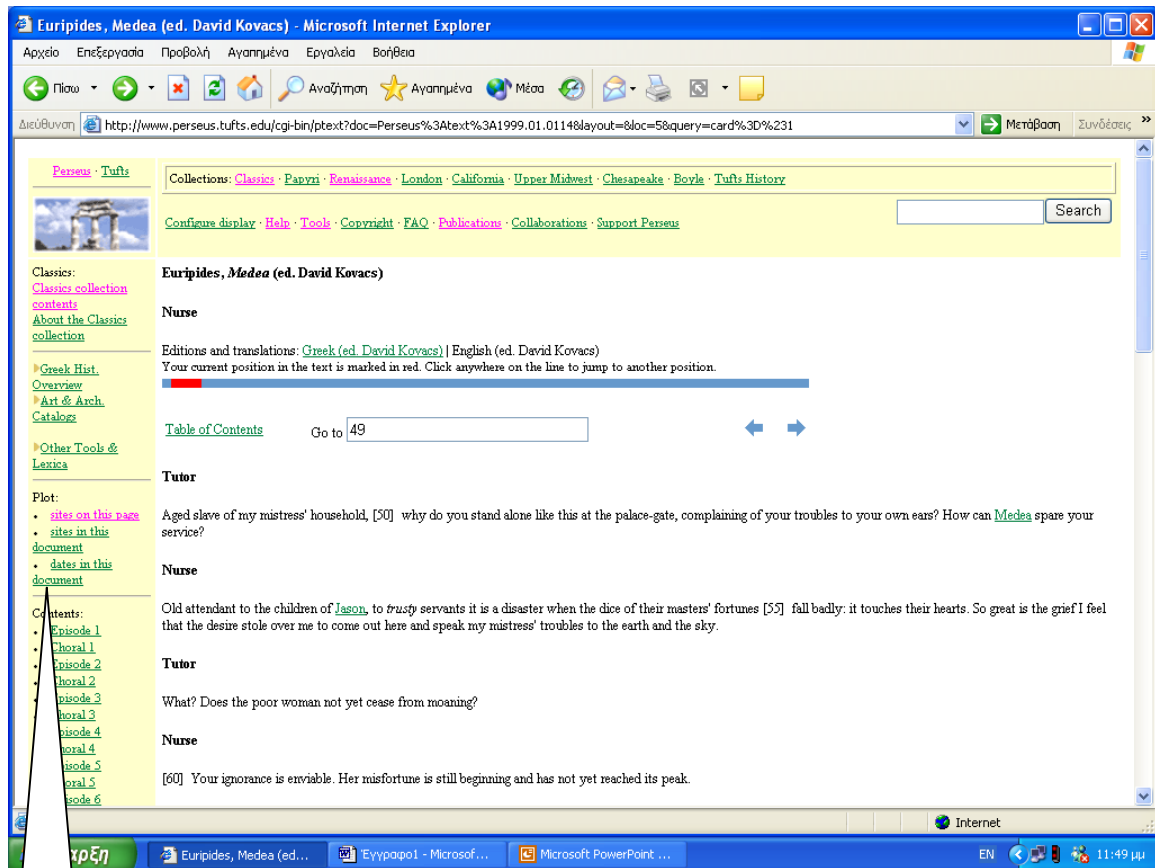
11.1 ΠΑΡΑΡΤΗΜΑ Α: Λεξιλόγιο όρων της αγγλικής και αποδόσεις στα ελληνικά

Aggregate=ομαδοποιώ
 Bounding rectangle=ορθογώνιο
 Centroid=κέντρο βάρους
 Clustering=συσταδοποίηση
 Context rules=κανόνες συμφραζόμενων
 Disambiguation=αποσαφήνιση
 Document context=συμφραζόμενα του τεκμηρίου
 Document Type Definition (DTD)=δήλωση τύπου εγγράφου
 Domain specific ontologies=οντολογίες συγκεκριμένου πεδίου ορισμού
 Explicit=ρητός
 Gateway=πύλη
 Gazetteer=γεωγραφικό λεξικό
 Geocoding=κωδικοποίηση γεωγραφικών δεδομένων
 Geographic interfaces=διεπαφές γεωγραφικών δεδομένων
 Geographic location=γεωγραφική θέση
 Geo-referenced digital library=ψηφιακή βιβλιοθήκη γεωγραφικών αναφορών
 Geospatial query=ερώτημα χωρικού τύπου
 Heuristic methods=ευρετικές μέθοδοι
 Honorifics=τιμητικές προσφωνήσεις
 Implicit georeferences=υπονοούμενες αναφορές γεωγραφικών χαρακτηριστικών
 Implicit searching=αναζήτηση υπονοούμενης πληροφορίας
 Integrated searching=ολοκληρωμένη αναζήτηση
 Interactive web interface=διαδραστική διεπαφή μέσω του Διαδικτύου
 Internal, external evidence=εσωτερικά, εξωτερικά στοιχεία
 Link=σύνδεσμος
 Local context=εντοπισμένα συμφραζόμενα
 Mapping application=εφαρμογή απεικόνισης
 Moving window= «κινούμενο παράθυρο»
 Named entity recognition=αναγνώριση ονομασμένης οντότητας
 Non-recursive entities=μη επαναλαμβανόμενες οντότητες
 One-by-one degree grid=πίνακας συντεταγμένων 1:1
 Place name=τοπωνύμιο
 Plot=σχεδιάζω, αποτυπώνω, παριστάνω γραφικά
 Probabilistic weight=βάρη πιθανοτήτων
 Scalable=διαβαθμισμένος
 Scan=σαρώνω
 Spatial references=αναφορές χώρων
 Standard deviation=τυπική απόκλιση
 String=σειρά
 Tag=κωδικοποιώ, επισημαίνω, ετικέτα
 Timeline=χρονολογική σειρά
 Training data=επεξεργασμένα δείγματα δεδομένων
 Variants=τύποι (λέξεων)
 Visualization=οπτικοποίηση
 Weighted map=πίνακας βαρών
 Word-occurrence pattern=πρότυπο παράλληλης εμφάνισης λέξεων

11.2 ΠΑΡΑΡΤΗΜΑ Β: Παράδειγμα αναζήτησης και αποτύπωσης χωρικής πληροφορίας στην Perseus Digital Library.

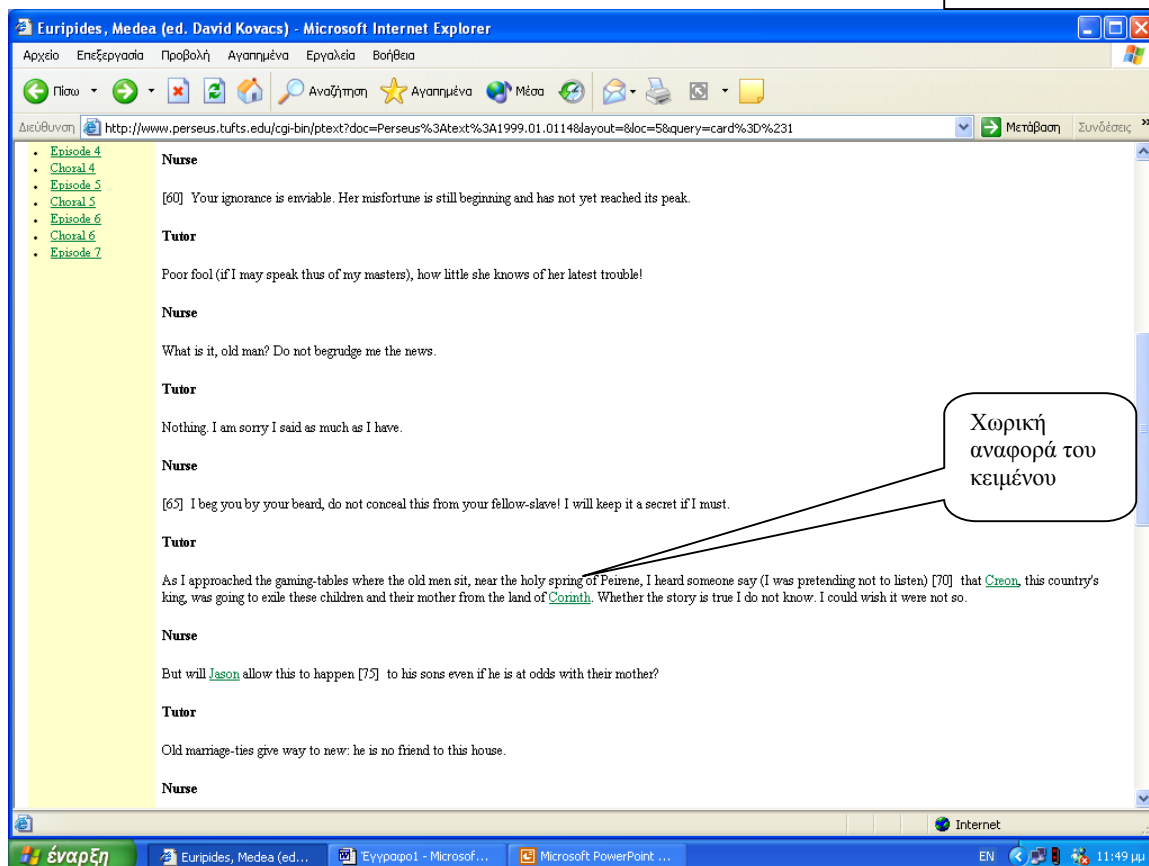
Στην Εικόνα 1 βλέπουμε τα αποτελέσματα μιας απλής αναζήτησης στην ψηφιακή βιβλιοθήκη για το έργο «Μήδεια» του Ευριπίδη. Βλέπουμε ένα τμήμα του έργου (αγγλική μετάφραση) και στην Εικόνα 2 τη συνέχεια της σελίδας όπου αναφέρεται (ανάμεσα σε άλλες λέξεις-κύρια ονόματα) η πόλη Κόρινθος η οποία είναι σύνδεσμος. Αν διαλέξουμε την επιλογή “Plot: Sites on this page” που βρίσκεται αριστερά στην Εικόνα 1 το σύστημα μπορεί να εμφανίσει το σημείο του χάρτη της Ελλάδας που βρίσκεται η Κόρινθος (Εικόνα 3).

Εικόνα 1

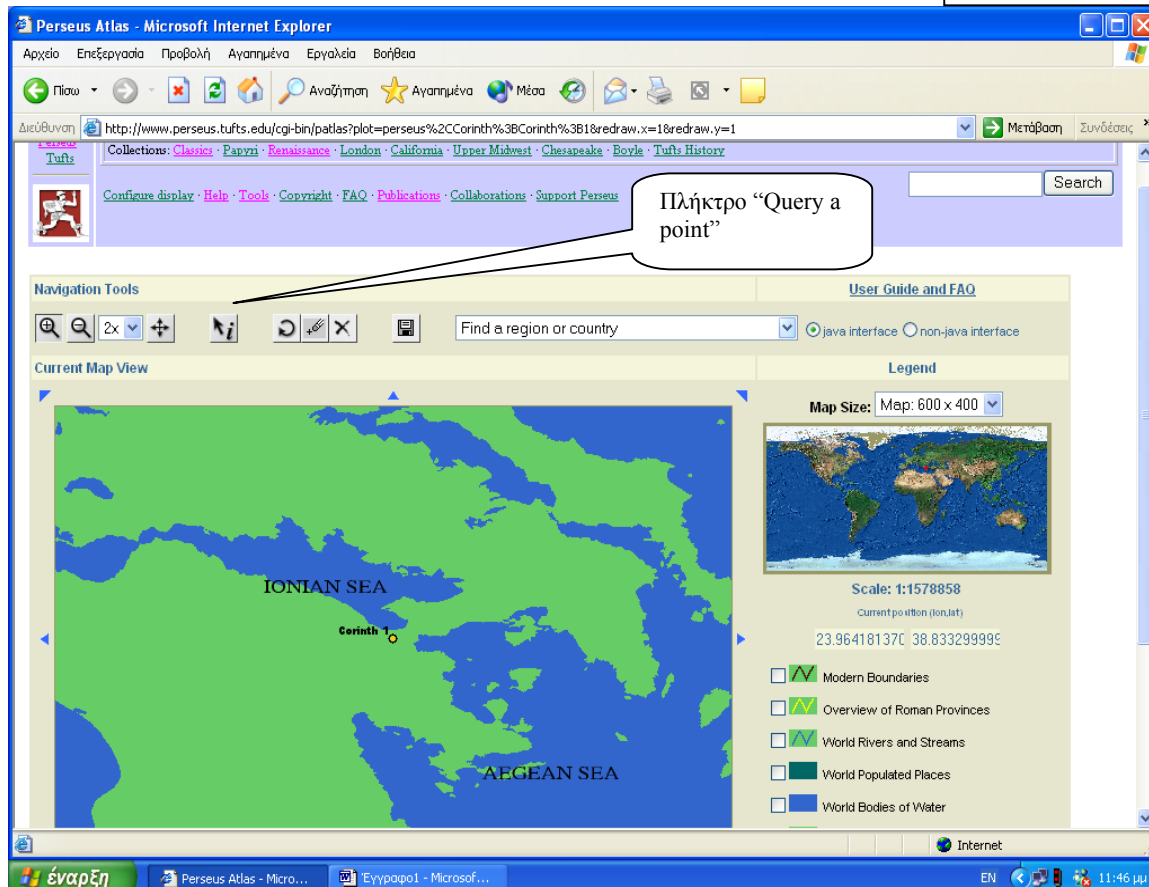


Δυνατότητα κατασκευής χαρτών και χρονολογικών σειρών

Εικόνα 2



Εικόνα 3



Στην Εικόνα 4 βλέπουμε τα αποτελέσματα της έρευνας με βάση την αναφορά του χάρτη. Επιλέγοντας το πλήκτρο του χάρτη “Query a point” (Εικόνα 3) μπορούμε να φέρουμε στην οθόνη όλες τις αναφορές για το σημείο του χάρτη που επιλέγουμε και υπάρχουν στην ψηφιακή βιβλιοθήκη. Επιλέγοντας το σημείο «Κόρινθος» βλέπουμε όλες τις αναφορές για την Κόρινθο που υπάρχουν στα τεκμήρια της ψηφιακής βιβλιοθήκης.

Εικόνα 4

Perseus Lookup Tool - Microsoft Internet Explorer

Αρχείο Επεξεργασία Προβολή Αγαπημένα Εργαλεία Βοήθεια

Πίσω Αναζήτηση Αγαπημένα Μέσα Μετάβαση Συνδέσεις

http://www.perseus.tufts.edu/cgi-bin/vor?type=phrase&lookup=Corinth

Perseus Tufts

Collections: [Classics](#) · [Papyri](#) · [Renaissance](#) · [London](#) · [California](#) · [Upper Midwest](#) · [Chesapeake](#) · [Boyle](#) · [Tufts History](#)

[Configure display](#) · [Help](#) · [Tools](#) · [Copyright](#) · [FAQ](#) · [Publications](#) · [Collaborations](#) · [Support Perseus](#)

Perseus Lookup Tool

Searched all Perseus collections for "Corinth"

Included alternate terms: Corinthos Ephyræa Korinth Korinthos

Results page: 1 2

60 from [Perseus Tools and Information](#)

1. [Corinth](#): Alabama, United States [Atlas site] (8.29)
2. [Corinth Corner](#): Vermont, United States [Atlas site] (8.07)
3. [Corinth](#): Illinois, United States [Atlas site] (7.84)

[More](#)

133 from [Greek and Roman Materials](#)

1. [CORINTH \(Korinthos\) Corinthia, Greece](#) (Reference article in *The Princeton Encyclopedia of Classical Sites* (eds. Marian Holland McAllister, Richard Stillwell, William L. MacDonald)) (13.92)
2. [KORINTHOS, see CORINTH, KORINTHOS, KORINTHOS, see CORINTH](#) (Reference article in *The Princeton Encyclopedia of Classical Sites* (eds. Marian Holland McAllister, Richard Stillwell, William L. MacDonald)) (13.85)

[Corinth, Corinth, Fountain of Glauke](#): Destroyed facade of fountainhouse from N [Image] (7.93)

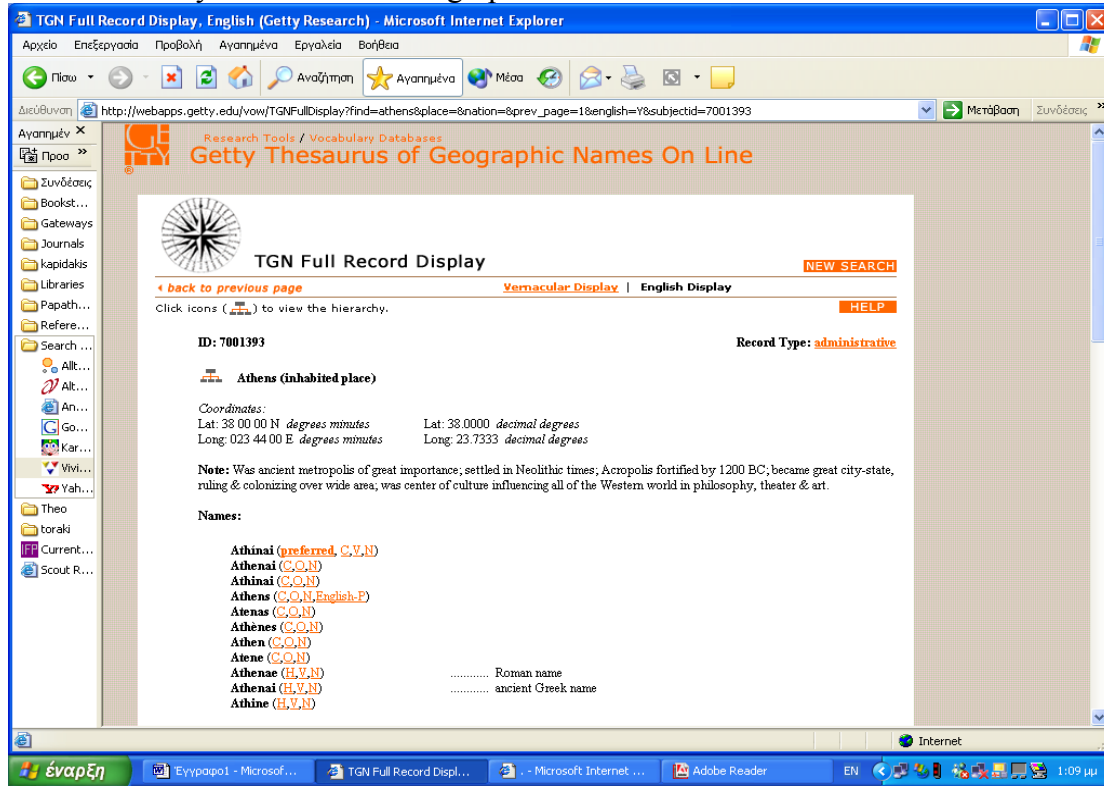
Results summary (items)

- [Perseus Tools and Information](#) (60)
- [Greek and Roman Materials](#) (133)
- [American Civil War](#) (389)
- [American Memory: Upper Midwest](#) (3)
- [Beazley Archive](#) (810)
- [CIMI Metadata Harvesting Working Group Demonstration...](#) (36)
- [Library of Congress Open Archive Initiative...](#) (3)
- [OCLC Online Computer Library Center Theses...](#) (7)
- [The American Numismatic Society](#) (2)
- [The University of Michigan University Library...](#) (3)
- [University of Illinois Library](#) (1)

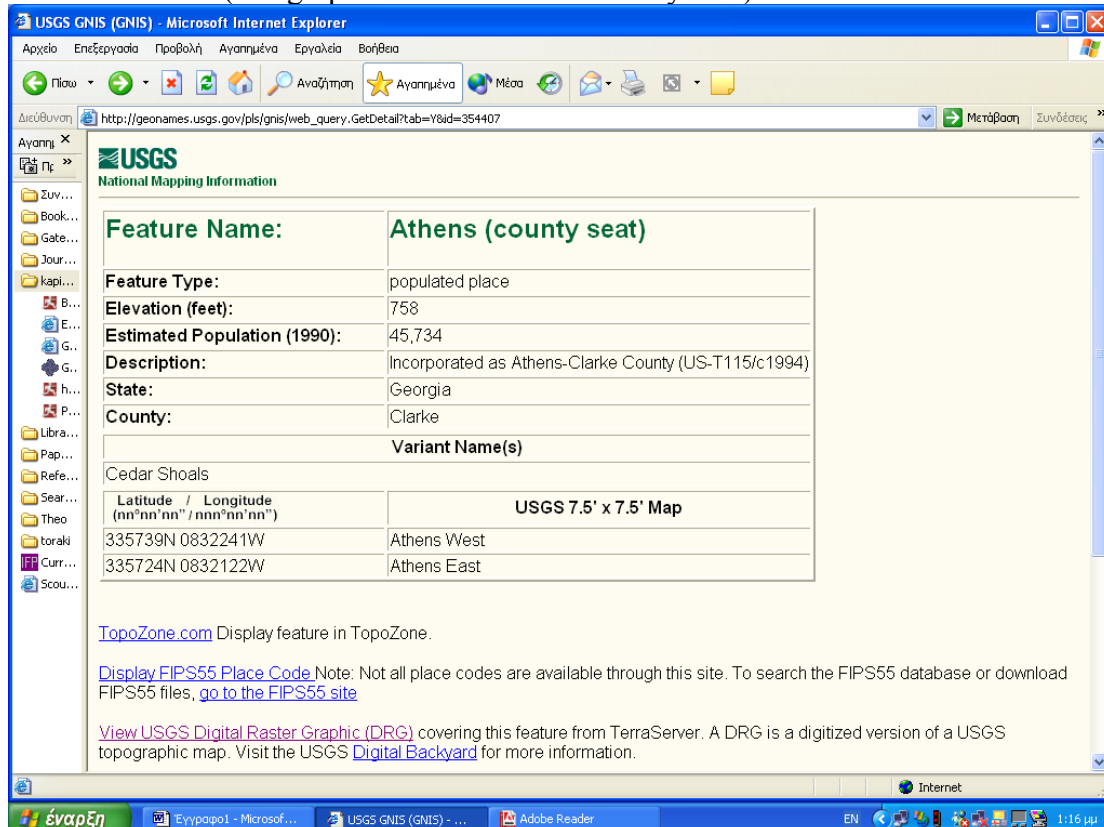
Σύνολο αναφορών για το σημείο του χάρτη που επιλέξαμε

11.3 ΠΑΡΑΡΤΗΜΑ Γ: Παραδείγματα εγγράφων ψηφιακών γεωγραφικών λεξικών

Εικόνα 1: Getty Thesaurus of Geographic Names on line



Εικόνα 2: GNIS (Geographic Names Information System)



11.4 ΠΑΡΑΡΤΗΜΑ Δ: Άρθρο αναφοράς.

Smith, D.A., Crane, G. (2001). Disambiguating geographic names in a historical digital library. In P. Constantopoulos and I.T. Solvberg (eds.): ECDL 2001, LNCS 2163, (pp.127-136), Berlin: Springer-Verlag, 2001. Available at: <http://www.perseus.tufts.edu/Articles/geodl01.pdf> (last accessed at 23 January 2004).

Disambiguating Geographic Names in a Historical Digital Library*

David A. Smith and Gregory Crane

Perseus Project
Tufts University
Medford, MA, USA
`{dasmith,gcrane}@perseus.tufts.edu`

Abstract. Geographic interfaces provide natural, scalable visualizations for many digital library collections, but the wide range of data in digital libraries presents some particular problems for identifying and disambiguating place names. We describe the toponym-disambiguation system in the Perseus digital library and evaluate its performance. Name categorization varies significantly among different types of documents, but toponym disambiguation performs at a high level of precision and recall with a gazetteer an order of magnitude larger than most other applications.

1 Introduction

Geographic interfaces provide natural, scalable visualizations for many digital library collections. Although domain-specific ontologies or automatic clusterings of documents may produce productive browsing tools in many cases, real world maps, along with timelines, can situate a wide range of information in a consistent, familiar space. When the contents of digital library documents are georeferenced, users can get a sense of the scope and focus points of a collection or a document, plot geographic places mentioned on any page of text, or find information about the places mentioned on a map or in a region [4,6]. At the Perseus Project, we have concentrated on representing historical data in the humanities from ancient Greece to nineteenth-century America [9]. With over one million identified toponym references, Perseus has built a rich digital library testbed and toolset that is available over the World Wide Web (<http://www.perseus.tufts.edu>; see Fig. 1, 2, and 3).

In order to reap the benefits of geographic interfaces, digital librarians must identify geographic names and link them to information about their location, in most cases their type (e.g. river, mountain, populated place), and other useful information such as dates of occupation, population at various times, and relation to other places. For documents of highly central importance to a scholar's work, it

* This research was supported by a grant from the Digital Libraries Initiative, Phase 2, with primary funding from the National Science Foundation and National Endowment for the Humanities.

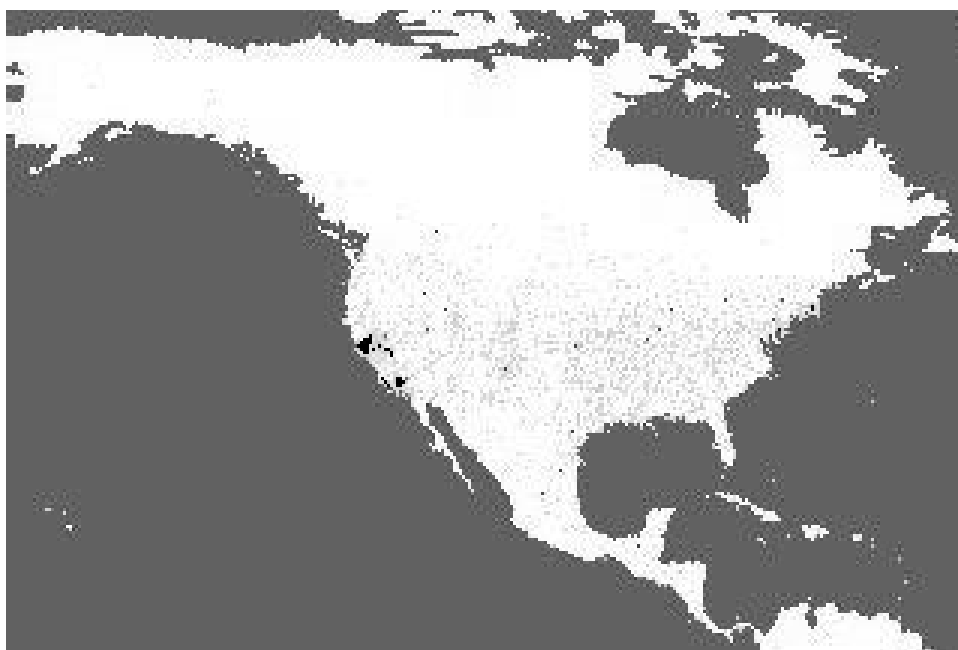


Fig. 1. The scope and focus of the collection on the settlement of California. Note the fainter spread of sites across the U.S. and the concentration in northern California.



Fig. 2. Sites mentioned in Herodotus. Note the strong concentration in present-day Greece and western Turkey.

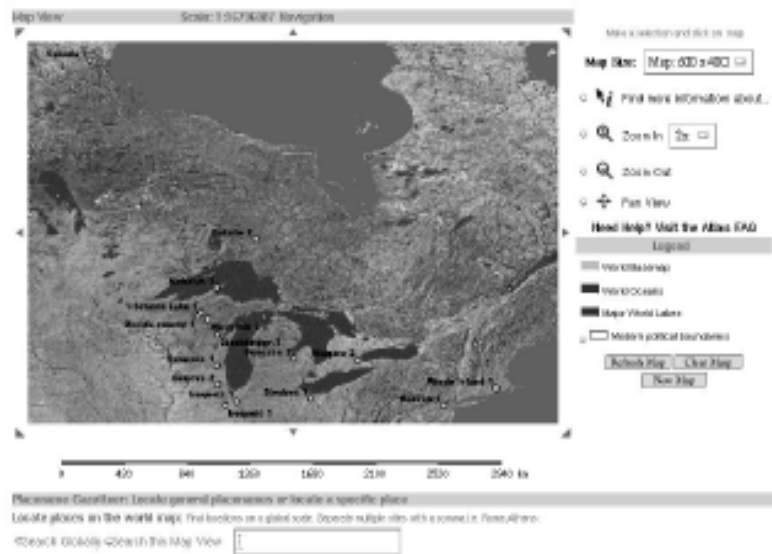


Fig. 3. Interactive map of the sites mentioned on one page of a diary of a voyage from Detroit to the source of the Mississippi, where the author is in Wisconsin but makes peripheral reference to the Narragansetts of Rhode Island. The user can zoom in on a particular region, such as the cluster in Oneida county, Wisconsin, or click on a site for more information.

might be worthwhile to spend the effort of manually tagging and disambiguating place names in a text, but manually tagging an entire corpus of any considerable size is impractical. Even at an optimistic ten seconds per toponym, it would take 28,000 person-hours to check the over one million toponyms in the Perseus DL of about 70 million words of English. We thus need automatic, or at least machine-assisted, methods for building georeferenced digital libraries.

2 Problem Description and Related Work

Linking strings in documents to locations on a map can be generally divided into two steps: name identification and categorization and disambiguation of those names classed as toponyms against a gazetteer with at least some coordinate information. In the past decade, many projects have devoted themselves to the first step and many fewer to the second. This concentration is not surprising; “named entity recognition”, as the name categorization task is known, aims to classify entities as persons, organizations, dates, products, organisms, and so on, in addition to geographic entities, and has many applications in the fields of message understanding and information extraction as a whole. Two general strategies for named entity recognition can be represented by two widely known systems. EBN’s Nymble [2] performs quite well with F-measures (see equation 1 below) at 90% or above, but requires at least 100,000 words of training data; IBM’s Nominator [10] performs only slightly less well (F-measure approximately 88%) on Wall Street Journal documents with only simple heuristics, though the authors admit that the “heuristics are somewhat domain dependent”.

The Geo-Referenced Information Processing System (GIPSY) described by [11] and [8] matches geographic names in text to spatial coordinates. Interestingly, this system also attempts to match such phrases as “south of Lake Tahoe” with fuzzy polygons. [8] describes a system to plot locations mentioned in transcripts of news broadcasts. Using a gazetteer of about 80,000 items, they report matching 289 out of 387 places (75%) in the test segments. Kanada reports 98% precision for geographic name disambiguation in Japanese text with a gazetteer of 88,000 Japanese and 41,000 foreign ones [8]. In an interesting parallel to our results below, he also reports significantly lower precision for Japanese toponyms than for foreign ones.

The documents in many digital libraries, however, present some particular problems for automatically identifying and disambiguating place names. Much work on proper names has dealt with news texts with useful discourse conventions for reducing ambiguity. A story mentioning Bill Clinton will use the full title “President Bill Clinton” on the first mention and “Mr. Clinton” or “Clinton” only afterwards. News stories also have relatively small scope, without long-distance anaphora. When a new story begins, President Bill Clinton is named in full all over again. Finally, place names themselves almost always have a disambiguating tag at their first mention, e.g. “London, Ontario” or “Clinton, New Jersey”. Digital libraries, on the other hand, often contain documents of widely varying lengths written without benefit of journalistic style. Scholarly works of

ten deal with several registers of time and place: a book on Shakespeare will talk not only about sixteenth-century Stratford (Warwickshire) but also about scholarship in nineteenth-century Cambridge (Massachusetts) and twentieth-century Berkeley (California). A wide historical purview can also make some pieces of knowledge in a gazetteer useless or misleading. Although the city of Samos is now in Greece and Miletus is in Turkey, they were both founded by Ionian Greeks and are only about 30 kilometers apart. Actual distance on the earth tells more than modern political categories.

Finally, a heterogeneous digital library can benefit from large knowledge bases to deal with its diverse materials but must deal with the cost of clashes among items in these authority lists. We can explore some of these ambiguities a priori by looking at the distributions in a gazetteer (table 1). Although the proportions are dependent on the names and places selected for inclusion in this gazetteer, the relative rankings are suggestive. In long-settled areas—such as Asia, Africa, and Europe—a place may be called by many names over time, but individual names are often distinct. With the increasing tempo of settlement in modern times, however, many places may be called by the same name, particularly by nostalgic colonists in the New World. Other ambiguities arise when people and places share names. Very few Greek and Latin place names are also personal names. This is less true of Britain, where surnames (and surnames used as given names) are often taken from place names; in America, the confusion grows as numerous towns are named after prominent or obscure people. In practice, we can express the scope of the disambiguation problem as follows: not counting the other names that could be mistaken for place names, some 92% of the toponyms in the Perseus digital library refer, potentially, to more than one place.

Continent	% places w/multiple names	% names w/multiple places
North & Central America	11.5	57.1
Oceania	6.9	29.2
South America	11.6	25.0
Asia	32.7	20.3
Africa	27.0	18.2
Europe	18.2	16.6

Table 1. Places with multiple names and names applied to more than one place in the Getty *Thesaurus of Geographic Names*

3 Disambiguation Procedure

As mentioned above, toponym disambiguation consists of name identification and categorization and disambiguation of those names. Our methods for performing these tasks rely on evidence that is internal or external to the text. (Note

the difference with the terminology and approach in [7], which uses evidence internal and external to the name.) Internal evidence includes the use of honorifics, generic geographic labels, or linguistic environment. External evidence includes gazetteers, biographical information, and general linguistic knowledge.

Before either identification or disambiguation could proceed, we gathered the knowledge sources used to make the categorization and disambiguation decisions. *Persius* uses some knowledge sources, such as the Getty Thesaurus of Geographic Names or Cruchley's gazetteer of London, that were purpose-built for geocoding. We captured other information, such as lists of authors or the entries in the *Dictionary of National Biography*, as a by-product of constructing the digital library as a whole. In total, the gazetteer used for name identification and disambiguation contains over one million place names.

We then scan the documents in the digital library for possible proper names and assign the names, if possible, to broad categories such as person, place, or date. We chose to use simple heuristic methods like those used in *Nominator* [10] rather than learning systems, since we lacked training data for our types of documents, and since we were mostly interested in identifying geographic names and not in the broader task of named entity recognition and categorization. In English text, the *Persius* system exploits generally used capitalization and punctuation conventions: initial candidates proper names are strings of capitalized words, and sentences are delimited with periods. Also at this stage, we exploit any markup that a document's editor has added, whether in tagging a string as a personal or place name, or in explicitly linking that name to an entry in a gazetteer. For initial categorization, the *Persius* system uses language-specific honorifics (such as "Dr." or "Mrs.") as strong evidence that the following name is a personal name. In addition, once a "Col. Aldrich", for example, is seen in a document, further references to "Aldrich" are automatically classified a personal names. Generic topographic labels (such as "Rocky Mountains" or "Charles River") are taken as moderate evidence that the name is geographic. Standalone instances of the most common given names in *Persius*' biographical dictionaries are labeled as personal names since a mere "John" is highly unlikely to refer to a town by that name in Louisiana or Virginia.

The system then attempts to match the names chosen as geographic, as well as the uncertain names, against a gazetteer. As our aim is to allow for geographic browsing of a digital library, it is of little benefit if we identify a name that cannot be linked to spatial coordinates. As mentioned above, for the names in our corpus that have at least one match in the gazetteer, about 92% match more than one entity.

Disambiguating the possible place names then proceeds based on local context, document context, and general world knowledge. The simplest instances of local context are the explicit disambiguating tags that authors put after place names: e.g. "Lancaster, PA", "Vienna, Austria", or "Beverly Hills, 90210". More generally, a place will more likely than not be near to other places mentioned around it. If "Philadelphia" and "Harrisburg" occur in the same paragraph, a reference to "Lancaster" is more likely to be to the town in Pennsylvania than

to the one in England or Arizona. Document context can be characterized as the preponderance of geographic references in an entire document; for short documents, such as news articles, local and document context can be treated as the same. World knowledge may be captured from gazetteers or other reference works and comprises such facts about a place as its coordinates, political entities to which it belongs, and its size or relative importance.

The system begins by producing a simple characterization of the document context. All of the possible locations for all of the toponyms in the document are aggregated onto a one-by-one degree grid, with weights assigned for the number of mentions of each toponym. The system prunes some possibilities based on general world knowledge, so that only Spain the country, and not the town in Tennessee, will be counted. We compute the centroid of this weighted map and the standard deviation of the distance of the points from this centroid. We then discard points more than two standard deviations away from the centroid and calculate a new centroid from the remaining points, if any.

We then process the possible toponyms for final disambiguation. We represent the local context of a toponym's occurrence as a moving window of the four toponyms mentioned before it and the four after it. Only unambiguous or already disambiguated toponyms are taken into account, however, in constructing this context. Each possible location for a toponym is given a score based on (a) its proximity to other toponyms around it, (b) its proximity to the centroid for the document, and (c) its relative importance—e.g. all other things being equal, nations get a higher score than cities. Also at this stage, the system discards as probable false positives places that lack an explicit disambiguator, that receive a low importance score, and that are far away from the local and document centroids. If not thus eliminated, the candidate toponym identification with the highest score is declared the winner. Once the work of the disambiguation system is done, the resulting toponyms are loaded into a relational database for access by the runtime digital library system.

4 Evaluation

We evaluate the performance of the disambiguation system using standard precision and recall methods. Qualitatively, the system performs quite well at producing aggregate maps of the places mentioned in an entire document or corpus or in finding mentions of a particular place. For a more detailed look at the performance of the *Perseus* system on various tests, a human disambiguator worked through 20% of the output for a text from each of five representative corpora: ancient Greece, ancient Rome, the Bodley collection on the history and topography of London, and two Library of Congress collections on the settlement of California and the Upper Midwest. With a large gazetteer and conservative pruning rules, our system is biased towards more recall. In table 2, we show precision results for the system as a whole, which is what the end user actually experiences, and for the toponym disambiguation system independent of name categorization. We also show the F-measure for the whole system, a score that

combines recall (R) and precision (P) with the recall/precision weighting factor β^2 usually valued at 1:

$$F = \frac{(\beta^2 + 1)RP}{(\beta^2 R) + P} \quad (1)$$

Corpus	Precision	Perfect Categ.	Recall	F-measure
Greek	0.93	0.98	0.99	0.96
Roman	0.91	0.99	1.00	0.95
London	0.86	0.92	0.96	0.91
California	0.83	0.92	0.96	0.89
Upper Midwest	0.74	0.89	0.89	0.81

Table 2. Performance on five representative texts

From these figures, one can see that although our simple heuristic categorization algorithm was less adequate for certain tasks, the toponym disambiguator itself performed quite well. The evaluation of toponym disambiguation is, if anything, conservative since eliminating extraneous points from the local and document context should reduce the skew in the calculated centroids. Note also that the categorization performed better on the Greek and Roman history texts than on texts on the history of London, California, or the Upper Midwest. This reflects the degree to which toponyms are ambiguous with other names or non-names in the text (see table 1 above). This evaluation also turned up another linguistic issue: all of the mistaken toponym identifications in the Roman text—Caesar’s *Gallie War*—were for the “Germans” whom Caesar is fighting. The ethnonym “German” is in the *gazetteer* in the record for Germany, but its plural is not. We could fill this deficiency by stemming the input, but proper names are not generally inflected in English, so on the whole stemming would do more harm than good. We can easily add these inflected geographic names to the *gazetteer* by hand. In general, the large *gazetteer* of over a million names probably depresses precision more than any other factor. In [6], for example, the *gazetteer* is an order of magnitude smaller (98,000) and precision reaches 98%.

5 Future Work

Although the Perseus toponym disambiguation system performs quite well, we will concentrate on improving the categorization system, especially for texts on North America. Many approaches to categorization require training data so that the system can learn context rules for the occurrence of various kinds of named entities. As noted above, important or canonical texts would in any case benefit from detailed hand markup, including name categorization and disambiguation,

and systems such as Almbic [3] have demonstrated computer-assisted methods to optimize the tagging task.

Restricting the available toponyms at any point in the text by time period would also improve the system's performance. In a heterogeneous digital library of historical information, however, a mix of temporal references may occur in close proximity. We could, however, use the preponderance of temporal references, as we now use the weighted map of spatial references, to rank the possibilities. "Ovid" in a discussion of Roman poetry is unlikely to refer to Ovid, Ithaca. While the current system deduces this from the town's distance from Italy, where most of the other places in the document are located, the fact that the town was founded in the nineteenth century would also tend to exclude it from a document where most of the dates are in the first centuries B.C. and A.D.

As explained above, we characterize the document context or central "region of interest" of a document by the centroid of the most heavily referenced areas. There seems to be some lack of robustness in simply using the centroid, and we are experimenting with using a bounding rectangle or polygon to represent a document's region of interest.

Finally, we are sampling on a gazetteer of Greek and Latin toponyms to apply this work to the non-English texts in the Perseus digital library. Much of this information can be culled from digitized reference works such as the *Harper's Dictionary of Classical Antiquities* and *Smith's Dictionary of Greek and Roman Geography*. Although we note above that morphological stemming of the source text would be counterproductive for English, we will need to stem Greek and Latin texts with our existing tools in order to perform well with these highly inflected languages.

References

1. Association for Computational Linguistics. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, April 1987.
2. Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pages 194–201.
3. David Day, John Abordas, Lynette Hirschman, Robyn Kentorok, Patricia Robinson, and Marc Vilain. Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pages 348–353.
4. Linda L. Hill, James Frow, and Qi Zheng. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1), January 1999. See <http://www.dlib.org/dlib/january99/hill/01hill.html>.
5. Yasuji Kanada. A method of geographical name extraction from Japanese text for thematic geographical search. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 46–54, Kansas City, Missouri, November 1998.
6. Ray R. Larson. Geographic information retrieval and spatial browsing. In Linda C. Smith and Myke Gluck, editors, *Geographic Information Systems and Libraries*.

- Patterns, Maps, and Spatial Information*, pages 81–123, April 1995. See http://inkarlock.berkeley.edu/gso_in/F&T1.html.
7. David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Bogunov and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 21–39. MIT Press, Cambridge, MA, 1996.
 8. Andreas M. Olligschlaeger and Alexander C. Hauptmann. Multimodal information systems and GIS: The Informedia digital video library. In *Proceedings of the ESRI User Conference*, San Diego, California, July 1999.
 9. David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.
 10. Nina Wacholder, Yoel Ravin, and Mitook Choi. Disambiguation of proper names in text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pages 302–308.
 11. Allison G. Woodruff and Christian Hauni. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655, 1994.