Ανάλυση χειρογράφων

Τα κείμενα των ανθρωπιστικών σπουδών είναι πολύπλοκα, μεγάλη ποικιλία δεδομένων και πλήθος διαφορετικών πληροφοριών σε διάφορες μορφές τα οποία πρέπει να εντοπιστούν και να περιγραφούν. Συγκεκριμένα:

- Περιλαμβάνουν μεγάλη ποικιλία υλικών γραφής και φυσικής εμφάνισης.
- Έχουν πολύπλοκη οργάνωση, δηλαδή στο κείμενο συνυπάρχουν πολλές και διαφορετικές δομές οι οποίες πρέπει να αποτυπωθούν με επικαλυπτόμενες ιεραρχίες σήμανσης. (Ποιήματα – π.χ. στίχους, στροφές ή ενότητες, Πεζά κείμενα - π.χ. ενότητες, κεφάλαια και παραγράφους, Θεατρικά / δραματικά κείμενα – π.χ. πράξεις, σκηνές, διαλόγους και μονολόγους, Προφορικό λόγο / ομιλία προσώπων – π.χ. ευθύς/πλάγιος λόγος, Ιδιαίτερο λεξιλόγιο - π.χ. ονόματα, ορολογίες, ημερομηνίες).

Επιπλέον, ο ερευνητής (αλλά και ο εκδότης / επιμελητής) συχνά καλείται να εντοπίσει, να αναγνωρίσει και να αποτυπώσει όλα τα παραπάνω και επιπλέον:

- Τα αλλεπάλληλα στάδια επεξεργασίας μέσα στο κείμενο.
- Τις διάφορες εκδοτικές και συγγραφικές επεμβάσεις στο πρωτότυπο.
- Τις ενδοκειμενικές και διακειμενικές αναφορές.
- Την ύπαρξη μη κειμενικού υλικού (π.χ. εικόνες), βιβλιογραφικές και εκδοτικές πληροφορίες αλλά και πληροφορίες που αφορούν στην εμφάνιση του κειμένου (π.χ. έντονη ή πλαγιογράμματη γραφή).

Όλα τα παραπάνω, καθιστούν την μελέτη και διερεύνηση των χειρογράφων είναι μια σύνθετη, απαιτητική και χρονοβόρα εργασία που απαιτεί εξειδικευμένα λογισμικά εξόρυξης και απεικόνισης όλων των χρήσιμων πληροφοριών και των δεδομένων ενός κειμένου τα οποία πρέπει να προσαρμοστούν σε συγκεκριμένες ερευνητικές ανάγκες. Απαραίτητη προϋπόθεση για την αξιοποίηση των παραπάνω λογισμικών είναι να προηγηθεί η ψηφιακή αναπαράσταση (κωδικοποίηση / encoding) του κειμένου.

Τα παραπάνω λογισμικά πρέπει να προσαρμοστούν σε συγκεκριμένες ερευνητικές ανάγκες και επομένως προϋποθέτουν (κατόπιν της διαδικασίας ψηφιακής μεταγραφής, παραγωγή ηλεκτρονικού κειμένου):

- Την ψηφιακή αναπαράσταση της δομής και των περιεχομένων των κειμένων τους
- Τον προγραμματισμό και πρόβλεψη για την κατάλληλη επιλογή της μεθοδολογίας κωδικοποίησης και σήμανσης αυξάνοντας σημαντικά το κόστος και την ανάγκη διασφάλισης της διαλειτουργικότητας και της μακροχρόνιας διατήρησης και διαθεσιμότητας.

Ψηφιακή αναπαράσταση κειμένου
(Text encoding)

Αναπαράσταση κειμένου είναι η διαδικασία κωδικοποίησης (σήμανσης / markup) των δομικών, μορφολογικών και σημασιολογικών χαρακτηριστικών του κειμένου.

Κατόπιν ακολουθεί η εισαγωγή του κωδικοποιημένου κειμένου σε βάση δεδομένων που θα επιτρέπει σύνθετες αναζητήσεις, εξόρυξη κειμένου (text mining) και ανάκτηση δεδομένων. Οι εργασίες σήμανσης αποδίδουν ένα κείμενο «εμπλουτισμένο» με πληροφορίες εκδοτικές, σημασιολογικές κ. ά., οι οποίες αποτελούν τη βάση των σύνθετων αναζητήσεων για τον χρήστη.

Στόχος είναι ο εμπλουτισμός του κειμένου με πληροφορίες οργανωμένες και ταξινομημένες σε δομημένα και διακριτά σύνολα.

Η πρόσβαση στα παραπάνω σύνολα δεδομένων του ηλεκτρονικού κειμένου επιτρέπει στον αναγνώστη, ερευνητή και εκδότη την αναζήτηση και ανάκτηση του κειμένου, καθώς επίσης και τη δυναμική διαχείριση, μελέτη, επεξεργασία, διαμόρφωση και αναπαράστασή του.

Το έργο της ψηφιακής αναπαράστασης (= σήμανσης) αποτελεί και αυτό φιλολογικό έργο, με ερμηνευτική διάσταση· δεν πρόκειται για αυτόματη τεχνική εργασία αλλά χειρωνακτική φιλολογική (ερμηνεία και σήμανση), που απαιτεί ωστόσο διεπιστημονική συνεργασία, πολλαπλά ωφέλιμη.

Τα στάδια

Αρχικά πρέπει να προσδιοριστεί το υποσύνολο των δεδομένων και των πληροφοριών του κειμένου που είναι χρήσιμα για τις ερευνητικές ανάγκες. Για παράδειγμα:

- Για τη μελέτη του λεξιλογίου ενός συγγραφέα είναι απαραίτητη η σήμανση και η διάκριση της γραφής του συγγραφέα απο παρένθετες γραφές που προέρχονται από άλλους συγγραφείς.
- Για τη μελέτη μιας ιδιολέκτου, μιας κοινωνιολέκτου ή ενός ιδιώματος σε μια συλλογή αρχείων είναι απαραίτητη η κατηγοριοποίηση του περιεχομένου και η ταξινόμησή του σε ειδικές λεξιλογικές ομάδες.
- Για τη μελέτη χειρόγραφων κειμένων είναι απαραίτητη η διάκριση και η ταυτοποίηση των επεμβάσεων και των αλλεπάλληλων σταδίων επεξεργασίας μέσα στο κείμενο.

Μετά ακολουθεί ο σχεδιασμός της μεθοδολογίας σήμανσης που αφορά στον προσδιορισμό των κανόνων που θα χρησιμοποιηθεί για την:

1. Δομική σήμανση του κειμένου, δηλαδή για την κωδικοποίηση της δομής και της ιεραρχίας των περιεχομένων του κειμένου.
2. Μορφολογική σήμανση του κειμένου, δηλαδή για την κωδικοποίηση της μορφής και της τυπογραφικής εικόνας του κειμένου.
3. Σημασιολογική σήμανση του κειμένου, δηλαδή για την κωδικοποίηση των παρακάτω στοιχείων του περιεχομένου του κειμένου:
   - Ενδοκειμενικά στοιχεία (π.χ. για τη συγκρότηση ευρετηρίου προσώπων και ξένων λέξεων).
   - Παρακειμενικά στοιχεία (αφιερώσεις, προμετωπίδες, τοποχρονολογίες, ομιλούντα πρόσωπα, σκηνικές οδηγίες κλπ.).
   - Εξωκειμενικά στοιχεία (συνοδευτικές σημειώσεις του ποιητή ή του εκδότη).

Η διαδικασία διάκρισης των δομικών, μορφολογικών και σημασιολογικών χαρακτηριστικών του κειμένου γίνεται με βάση κάποιους κανόνες (πρότυπο, σχήμα) και χρησιμοποιώντας ετικέτες (tags) οι οποίες «χαρακτηρίζουν» το κείμενο.

Κατόπιν ακολουθεί η εισαγωγή του κωδικοποιημένου κειμένου σε βάση δεδομένων που θα επιτρέπει σύνθετες αναζητήσεις, εξόρυξη κειμένου (text mining) και ανάκτηση και οπτικοποίηση δεδομένων.

Οι προκλήσεις

Οι βασικές προκλήσεις της κωδικοποίησης κειμένων είναι:

- Η εξασφάλιση ομοιομορφίας (στον τρόπο περιγραφή και αναπαράστασης των ηλεκτρονικών κειμένων) και διαλειτουργικότητας (δηλαδή η σήμανση των δομικών, μορφολογικών και σημασιολογικών χαρακτηριστικών και στοιχείων του κειμένου να αναγνωρίζεται από όλους τους υπολογιστές ώστε να εξασφαλιστεί / βελτιωθεί η αναζήτηση και η ανάκτηση της πληροφορίας που περιέχεται σε αυτά.)
- Η επίτευξη ισορροπίας μεταξύ μορφής και οπτικοποίησης του κειμένου
- Η μέριμνα για την ενσωμάτωση εργαλείων οπτικοποίησης και διαχείρισης που καθιστούν εφικτά τα παρακάτω:
    - Στατιστικής ανάλυσης (Λεξικογραφία και υφολογία).
    - χρηστική ή κριτική έκδοση, όπου θα είναι εφικτή η παράλληλη προβολή και αντιπαραβολή διαφορετικών εκδοχών του ίδιου κειμένου

Πληθώρα εργαλείων είναι διαθέσιμα στην σελίδα
https://digitalresearchtools.pbworks.com/w/page/17801708/Text-Analysis-Tools η οποία περιλαμβάνει και εργαλεία transcription και visualization (βλ αρχική σελίδα)

Για την αντιμετώπιση των παραπάνω προκλήσεων, καθώς και για την κάλυψη των αυξημένων και εξειδικευμένων αναγκών αποθήκευσης, ανταλλαγής, πρόσβασης, αναζήτησης, ανάκτησης, επεξεργασίας και έκδοσης των (λόγω πολυπλοκότητας) κειμένων των Ανθρωπιστικών σπουδών για την ανάδειξη, διαχείριση και επιστημονική επεξεργασία τους σε ηλεκτρονική μορφή, οδήγησε τον κλάδο της Πληροφορικής σε προσπάθειες αξιοποίησης των γλωσσών σήμανσης για την δημιουργία προτύπων περιγραφής και εργαλείων διαχείρισης των δεδομένων και μεταδεδομένων, των δομικών, μορφολογικών και σημασιολογικών χαρακτηριστικών του κειμένου.

Η κωδικοποίηση του ηλεκτρονικού κειμένου να γίνεται με βάση τους κανόνες και τις οδηγίες ενός προτύπου προκειμένου να εξασφαλιστεί ένας ομοιόμορφος τρόπος περιγραφής και αναπαράστασης των ηλεκτρονικών κειμένων και η σήμανση των δομικών, μορφολογικών και σημασιολογικών χαρακτηριστικών και στοιχείων του κειμένου να αναγνωρίζεται από όλους τους υπολογιστές.
Η κωδικοποίηση του ηλεκτρονικού κειμένου γίνεται με βάση:

- Μια γλώσσα σήμανσης, όπως π.χ. οι xml, xtml, xstl.

- Ένα πρότυπο και πρωτόκολλο, όπως το Text Encoding Initiative (TEI).

Η επεκτάσιμη γλώσσα επισημείωσης XML (eXtensive Markup Language)

Η XML αποτελεί ένα σύνολο κανόνων και συμβάσεων σήμανσης που χρησιμοποιείται για την οργάνωση, αναπαράσταση και διαχείριση.

Η XML σήμανση διαθέτει καθορισμένη και αυστηρή δενδροειδής δομή που επιτυγχάνεται με την χρήση ετικετών (tags), η οποία είναι ιδανική για να την αναπαράσταση της δομής και των συστατικών κάθε είδους κειμένου, καθιστώντας εφικτή τη δυναμική διαχείρισή του από τον αναγνώστη και τον μελετητή.

Η γλώσσα XML έχει συνδεθεί στενά με την τεχνολογία των συστημάτων βάσεων δεδομένων.

Ένα κείμενο XML αποτελεί μια συλλογή συσχετιζόμενων δεδομένων και συχνά αντιμετωπίζεται σαν μια Βάση Δεδομένων.

Παράδειγμα της δενδροειδούς δομής της XML

Στην δενδρική αναπαράσταση της δομής για την περιγραφή του μαθήματος (βλ παρουσίαση):

- Η ρίζα είναι το στοιχείο Μάθημα.
- Τα φύλλα (leaves) του δένδρου είναι τα στοιχεία: Τίτλος_Μαθήματος, Κωδικός, Συγγράμματα, Τίτλος_Συγγράμματος, ISBN.
- Τα κλαδιά του δένδρου (branches) είναι τα στοιχεία Τίτλος_Συγγράμματος, ISBN.

Το XML αρχείο του οποίου η δομή ακολουθεί την παραπάνω δενδρική αναπαράσταση.

Το πρότυπο κωδικοποίησης κειμένων TEI

(Text Encoding Initiative)

Το TEI αποτελεί μια από τις σημαντικότερες κειμενοκεντρικές εφαρμογές της ψηφιακής τεχνολογίας σήμανσης στο πεδίο των DH με βάση την γλώσσα XML. Είναι ένα διεθνώς αναγνωρισμένο και ανοιχτό πρότυπο που χρησιμοποιείται ευρέως από την επιστημονική κοινότητα των DH για την σήμανση / κωδικοποίηση κάθε είδους κειμένων (λογοτεχνικών, γλωσσολογικών και ιστορικών). Το TEI εξασφαλιστεί τη συμβατότητά τους με τους ηλεκτρονικούς υπολογιστές (διαλειτουργικότητα), δηλαδή έναν ομοιόμορφο τρόπο περιγραφής και αναπαράστασης των ηλεκτρονικών κειμένων που αναγνωρίζεται από όλους τους υπολογιστές.

Το TEI βασίζεται στην XML και επομένως μπορεί να αξιοποιήσει τις τεχνικές αποθήκευσης, αναζήτησης και ανάκτησης, διαχείρισης και εμφάνισης κωδικοποιημένων δεδομένων που αναπτύσσονται για τις βάσεις δεδομένων. Η δενδροειδής δομή της γλώσσας XML χρησιμοποιείται για να εκφράσει και να αναπαραστήσει την εσωτερική δομή των κειμένων. Ωστόσο, η δομή κάθε κειμένου ποικίλει ανάλογα με το είδος του. Για παράδειγμα, ένα δραματικό κείμενο αναλύεται σε πράξεις, σκηνές, διαλόγους και μονολόγους, ενώ ένα πεζό κείμενο σε ενότητες, κεφάλαια και παραγράφους. Αντίθετα, ένα ποίημα αποτελείται από στίχους, οι οποίοι μπορεί να σχηματίζουν στροφές ή ενότητες. Το πρότυπο κωδικοποίησης TEI προβλέπει τη δομική ποικιλία ανάμεσα στα κείμενα, καθώς προτείνει μια σειρά από διαφορετικά στοιχεία (elements) για την αναπαράσταση της διαίρεσης των κειμένων σε ενότητες, πράξεις, κεφάλαια, στροφές και στροφικά σύνολα. Επιπλέον, παρέχει στοιχεία που κωδικοποιούν την ομιλία των προσώπων (ευθύς/πλάγιος λόγος), το ιδιαίτερο λεξιλόγιο (π.χ. ονόματα, ορολογίες, ημερομηνίες), τις εκδοτικές και συγγραφικές επεμβάσεις στο πρωτότυπο, τις ενδοκειμενικές και διακειμενικές αναφορές, την ύπαρξη μη κειμενικού υλικού (π.χ. εικόνες) και, τέλος, την εμφάνιση του κειμένου (π.χ. έντονη ή πλαγιογράμματη γραφή). Επιπλέον, προδιαγράφει για κάθε XML έγγραφο μια κοινή ενότητα για την κωδικοποίηση βιβλιογραφικών πληροφοριών που αφορούν το κείμενο, αλλά και το εκάστοτε εγχείρημα της ψηφιακής έκδοσης. Κάθε σύνολο δεδομένων του κειμένου αναλύεται στα επιμέρους τμήματα που το συγκροτούν. Κάθε κείμενο μπορεί να διαθέτει έναν κατάλογο περιεχομένων, έναν πρόλογο, εισαγωγικά σημειώματα, αλλά και να διαιρείται σε κεφάλαια ή ενότητες, και αυτά έπειτα σε παραγράφους. Για την αναλυτική

κωδικοποίηση της διαίρεσης ενός κειμένου το TEI παρέχει μια σειρά από στοιχεία σήμανσης που αναπαριστούν τα συστατικά της ενδοκειμενικής δομής. Συγκεκριμένα, η διαίρεση με βάση το πρότυπο κωδικοποίησης TEI μπορεί να ξεκινά από τις ευρύτερες ενότητες ενός κειμένου, να προχωρά σε παραγράφους, περιόδους, φράσεις, λέξεις και να καταλήγει σε μεμονωμένους χαρακτήρες φωνημάτων.

Το TEI αποτελεί κατά κύριο λόγο περιγραφικό πρότυπο κωδικοποίησης δεδομένων, αλλά χαρακτηρίζεται και ως διαχειριστικό και δομικό πρότυπο:

- Περιγραφικό πρότυπο διότι περιγράφει πηγές με σκοπό την ανάκτηση και την ταυτοποίηση της πληροφορίας.
- Διαχειριστικό διότι παρέχει πληροφορίες για τη διαχείριση μίας πηγής όπως πότε και από ποιον έχει δημιουργηθεί τεχνικές λεπτομέρειες, πληροφορίες πρόσβασης κτλ
- Δομικό διότι υποδεικνύει τον τρόπο συνδυασμού των αντικειμένων ενός τεκμηρίου π.χ. τη διάταξη των σελίδων προκειμένου να διαμορφώσουν κεφάλαια.

Οι κατηγορίες σήμανσης των κειμένων με βάση το TEI είναι οι εξής:

- Σήμανση της δομής.
- Σήμανση της μορφής.
- Σήμανση του περιεχομένου (κυρίως σημασιολογική/semantic) π.χ.
  - ενδοκειμενικά στοιχεία (π. χ. για τη συγκρότηση ευρετηρίου προσώπων και ξένων λέξεων),
  - παρακειμενικά στοιχεία (αφιερώσεις, motti, προμετωπίδες, τοποχρονολογίες, ομιλούντα πρόσωπα, σκηνικές οδηγίες κλπ.),
  - εξωκειμενικά στοιχεία (συνοδευτικές σημειώσεις του ποιητή ή του εκδότη).

TEI Ενότητες *(Modules)*

Το TEI διαθέτει 21 modules και πάνω από 1500 δυνατούς συνδυασμούς στοιχείων (elements) και ιδιοτήτων (attributes), που μπορούν να χρησιμοποιηθούν για την περιγραφή και αναπαράσταση κάθε είδους κειμένου σε ψηφιακή μορφή.

Επίσης, κάθε στοιχείο μπορεί να ανήκει ως υποστοιχείο (φωλισμένο) σε ένα ή περισσότερα στοιχεία, δηλαδή σε μια ή περισσότερες κλάσεις (ομάδες υποστοιχείων που είναι φωλιασμένα σε ένα άλλο στοιχείο που ονομάζουμε ρίζα). Ανάλογα με τις κλάσεις που ανήκει ένα στοιχείο, το περιεχόμενο και τα γνωρίσματά του μπορεί να είναι διαφορετικά. Οι κλάσεις που ανήκει κάθε στοιχείο, το περιεχόμενο και τα γνωρίσματα του περιγράφονται πλήρως στον ορισμό του. Ο δισδιάστατος χώρος ενοτήτων – κλάσεων των στοιχείων προσφέρει ιδιαίτερη δυναμική και ευελιξία στο πρότυπο.

Κάθε στοιχείο ανήκει, επίσης, σε μια ή περισσότερες κλάσεις. Ανάλογα με τις κλάσεις που ανήκει ένα στοιχείο, το περιεχόμενο και τα γνωρίσματά του μπορεί να είναι διαφορετικά. Οι κλάσεις που ανήκει κάθε στοιχείο, το περιεχόμενο και τα γνωρίσματα του περιγράφονται πλήρως στον ορισμό του. Ο δισδιάστατος χώρος ενοτήτων – κλάσεων των στοιχείων προσφέρει ιδιαίτερη δυναμική και ευελιξία στο πρότυπο.

Για τον έλεγχο της εγκυρότητας των XML εγγράφων απαιτείται ένα μοντέλο εγγράφου το οποίο ορίζει τη σύνταξη ενός XML αρχείου καθώς αποτελεί ένα σύνολο κανόνων που παρέχουν το βασικό λεξιλόγιο και τη γραμματική για τον καθορισμό με ακρίβεια των στοιχείων που επιθυμούμε να θεωρούμε ως «έγκυρα», του περιεχομένου τους και των ιδιοτήτων τους και των σχέσεων μεταξύ τους, όπως και τη σειρά, αλλά και το πλήθος των εμφανίσεών τους.

Ως μοντέλο εγγράφου μπορεί να χρησιμοποιηθεί είτε μια Δήλωση Τύπου Αρχείων DTD (Document Type Definition) είτε ένα XML Σχήμα (Schema).

Ένα TEI σχήμα μπορεί να δομηθεί από ένα συνδυασμό ενοτήτων. Οι ενότητες μπορούν να συνδυάζονται και να σχηματίζουν TEI σχήματα που καλύπτουν συγκεκριμένες ανάγκες κωδικοποίησης. Κατά τον ορισμό ενός TEI σχήματος, το πρότυπο παρέχει τη δυνατότητα να προστεθούν στοιχεία σε κάποιες κλάσεις ή να αποκλειστούν κάποια άλλα στοιχεία από συγκεκριμένες ενότητες. Κάποιες ενότητες αποτελούν απαραίτητα συστατικά στις περισσότερες περιπτώσεις δόμησης σχημάτων.

Ενότητα		Περιγραφή
1	analysis	Analysis and Interpretation
2	certainty	Certainty and Uncertainty
3	core	Common Core
4	corpus	Metadata for Language Corpora
5	dictionaries	Print Dictionaries
6	drama	Performance Texts
7	figures	Tables, Formulae, Figures Documentation
8	gaiji	Character and Glyph
9	header	Common Metadata
10	iso-fs	Feature Structures
11	link	Linking, Segmentation, and Alignment
12	msdescription	Manuscript Description
13	namesdates	Names, Dates, People, and Places
14	nets	Graphs, Networks, and Trees
15	spoken	Transcibed speech
16	tagdocs	Documentation Elements
17	tei	TEI Infrastructure
18	textcrit	Text Criticism
19	textstructure	Default Text Structure
20	transcr	Transcription of Primary Sources
21	verse	Verse

Ενότητα	Περιγραφή - περιλαμβάνει στοιχεία
analysis	Τη γλωσσολογική ανάλυση των κειμένων (π.χ. για τη δήλωση προτάσεων, λέξεων, φράσεων, φωνημάτων κλπ)
certainty	Για τη δήλωση της αβεβαιότητας με την οποία γίνεται μια κωδικοποίηση ενός κειμένου (πχ μπορεί να μην είναι βέβαιο κάποιο μέρος ενός κειμένου (να έχουν χαθεί λέξεις, προτάσεις κλπ) ή να μην είναι απόλυτα βέβαιη η πνευματική ιδιοκτησία (ο συγγραφέας) ενός κειμένου)
core	που υπάρχουν σε όλα τα TEI τεκμήρια όπως δηλώσεις ονομάτων, ημερομηνιών, διευθύνσεων, σημειώσεις, συντομογραφίες, λέξεις από ξένες γλώσσες κλπ
corpus	μεταδεδομένων και κωδικοποίησης σωμάτων κειμένων (language copropa). Ένα σώμα κειμένου περιλαμβάνει συλλογές κειμένων ή τμημάτων τους με συγκεκριμένο γλωσσολογικό ενδαφέρον π.χ. κείμενα σε μια διάλεκτο μιας γλώσσας, ή κείμενα που αναπαριστούν όλες τις εκδοχές μιας γλώσσας
dictionaries	για έντυπα λεξικά
drama	που χρησιμοποιούνται για την κωδικοποίηση κειμένων που αφορούν σε παραστάσεις (θεατρο, μουσικόδραμα, κα)
figures	για πίνακες, σχήματα και μαθηματικούς τύπους που υπάχουν σε κείμενα
gaiji	για την τεκμηρίωση χαρακτήρων και των γραμματοσειρών τους
header	για τα μεταδεδομένα κωδικοποίησης ενός κειμένου
iso--fs	για την κωδικοποίηση καταλόγων χαρακτηριστικών με τις αντίστοιχες τιμές τους, πχ οι χώρες της Ευρώπης με τις πρωτεύουσές τους
link	για παραπομπές, για τη δόμηση του κειμένου σε επιμέρους τμήματα και τη συσχέτιση τμημάτων ή ολόκληρων κειμένων
msdescription	για την κωδικοποίηση χειρογράφων
namesdates	για την κωδικοποίηση ονομάτων, προσώπων και τοπωνυμίων
nets	Για την κωδικοποίηση γραφημάτων, δικτύων και διαγραμμάτων
spoken	για τη μεταγραφή προφορικού λόγου
tagdocs	για την τεκμηρίωση των αλλαγών που γίνονται στη σημασία ή τη συντακτική δομή των στοιχείων του TEI κατά τη δημιουργία TEI--σχημάτων

tei	περιλαμβάνει δηλώσεις και ορισμούς που αφορούν στη βασική υποδομή του προτύπου, όπως δηλώσεις τύπων δεδομένων, κλάσεων γνωρισμάτων κλπ
textcrit	που συνοδεύουν ένα γραπτό κείμενο, κριτικά υπομνήματα, αναλύσεις, πρώιμες μεταφράσεις κλπ
textstructure	που προδιαγράφουν τη βασική δομή ενός TEI τεκμηρίου και κωδικοποιούν τη δομή των κειμένων όπως στοιχεία για παράγραφους, κεφάλαια, κλπ
transcr	για την κωδικοποίηση του υποστρώματος στο οποίο υπάρχει ένα κείμενο ή χειρόγραφο, πχ μιας τηλεμοιοτυπίας (fax) που περιλαμβάνει ένα κείμενο ή ένα χειρόγραφο
verse	για την κωδικοποίηση του ποιητικού λόγου

Ενότητα Manuscript Description

Περιλαμβάνει τα παρακάτω 64 στοιχεία για την κωδικοποίηση χειρογράφων:

Name	Description
<accMat>	contains details of any significant additional material which may be closely associated with the manuscript being described, such as non-contemporaneous documents or fragments bound in with the manuscript at some earlier historical period.
<acquisition>	contains any descriptive or other information concerning the process by which a manuscript or manuscript part entered the holding institution.
<additional>	groups additional information, combining bibliographic information about a manuscript, or surrogate copies of it with curatorial or administrative information.
<additions>	contains a description of any significant additions found within a manuscript, such as marginalia or other annotations.
<adminInfo>	contains information about the present custody and availability of the manuscript, and also about the record description itself.
<altIdentifier>	contains an alternative or former structured identifier used for a manuscript, such as a former catalogue number.
<binding>	contains a description of one binding, i.e. type of covering, boards, etc. applied to a manuscript
<bindingDesc>	describes the present and former bindings of a manuscript, either as a series of paragraphs or as a series of distinct binding elements, one for each binding of the manuscript.
<catchwords>	describes the system used to ensure correct ordering of the quires making up a codex or incunable, typically by means of annotations at the foot of the page.
<collation>	contains a description of how the leaves or bifolia are physically arranged.
<collection>	contains the name of a collection of manuscripts, not necessarily located within a single repository.
<colophon>	contains the colophon of a manuscript item: that is, a statement providing information regarding the date, place, agency, or reason for production of the manuscript.
<condition>	contains a description of the physical condition of the manuscript.
<custEvent>	describes a single event during the custodial history of a manuscript.
<custodialHist>	contains a description of a manuscript's custodial history, either as running prose or as a series of dated custodial events.

<decoDesc>	contains a description of the decoration of a manuscript, either as a sequence of paragraphs, or as a sequence of topically organised decoNote elements.
<decoNote>	contains a note describing either a decorative component of a manuscript, or a fairly homogenous class of such components.
<depth>	specifies a length measured across the spine.
<dimensions>	contains a dimensional specification.
<explicit>	contains the explicit of a manuscript item, that is, the closing words of the text proper, exclusive of any rubric or colophon which might follow it.
<filiation>	contains information concerning the manuscript's filiation, i.e. its relationship to other surviving manuscripts of the same text, its protographs, antigraphs and apographs.
<finalRubric>	contains the string of words that denotes the end of a text division, often with an assertion as to its author and title, usually set off from the text itself by red ink, by a different size or type of script, or by some other such visual device.
<foliation>	describes the numbering system or systems used to count the leaves or pages in a codex.
<handDesc>	contains a description of all the different kinds of writing used in a manuscript.
<handNote>	describes a particular style or hand distinguished within a manuscript.
<height>	contains a measurement measured along the axis parallel to the spine.
<heraldry>	contains a heraldic formula or phrase, typically found as part of a blazon, coat of arms, etc.
<history>	groups elements describing the full history of a manuscript or manuscript part.
<incipit>	contains the incipit of a manuscript item, that is the opening words of the text proper, exclusive of any rubric which might precede it, of sufficient length to identify the work uniquely; such incipts were, in fomer times, frequently used a means of reference to a work, in place of a title.
<institution>	contains the name of an organization such as a university or library, with which a manuscript is identified, generally its holding institution.
<layout>	describes how text is laid out on the page, including information about any ruling, pricking, or other evidence of page-preparation techniques.
<layoutDesc>	collects the set of layout descriptions applicable to a manuscript.
<locus>	defines a location within a manuscript or manuscript part, usually as a (possibly discontinuous) sequence of folio references.

<material>	contains a word or phrase describing the material of which a manuscript (or part of a manuscript) is composed.
<msContents>	describes the intellectual content of a manuscript or manuscript part, either as a series of paragraphs or as a series of structured manuscript items.
<msDesc>	contains a description of a single identifiable manuscript.
<msIdentifier>	contains the information required to identify the manuscript being described.
<msItem>	describes an individual work or item within the intellectual content of a manuscript or manuscript part.
<msItemStruct>	contains a structured description for an individual work or item within the intellectual content of a manuscript or manuscript part.
<msName>	contains any form of unstructured alternative name used for a manuscript, such as an ocellus nominum, or nickname.
<msPart>	contains information about an originally distinct manuscript or part of a manuscript, now forming part of a composite manuscript.
<musicNotation>	contains description of type of musical notation.
<objectDesc>	contains a description of the physical components making up the object which is being described.
<origDate>	contains any form of date, used to identify the date of origin for a manuscript or manuscript part.
<origPlace>	contains any form of place name, used to identify the place of origin for a manuscript or manuscript part.
<origin>	contains any descriptive or other information concerning the origin of a manuscript or manuscript part.
<physDesc>	contains a full physical description of a manuscript or manuscript part, optionally subdivided using more specialised elements from the model.physDescPart class.
<provenance>	contains any descriptive or other information concerning a single identifiable episode during the history of a manuscript or manuscript part, after its creation but before its acquisition.
<recordHist>	provides information about the source and revision status of the parent manuscript description itself.
<repository>	contains the name of a repository within which manuscripts are stored, possibly forming part of an institution.
<rubric>	contains the text of any rubric or heading attached to a particular manuscript item, that is, a string of words through which a manuscript signals the beginning of a text division, often with an assertion as to its author and title, which is in some way set off from the text itself, usually in red ink, or by use of different size or type of script, or some other such visual device.
<seal>	contains a description of one seal or similar attachment applied to a manuscript

<sealDesc>	describes the seals or other external items attached to a manuscript, either as a series of paragraphs or as a series of distinct seal elements, possibly with additional decoNotes.
<secFol>	The word or words taken from a fixed point in a codex (typically the beginning of the second leaf) in order to provide a unique identifier for it.
<signatures>	contains discussion of the leaf or quire signatures found within a codex.
<source>	describes the original source for the information contained with a manuscript description.
<stamp>	contains a word or phrase describing a stamp or similar device.
<summary>	contains a brief summary of the intellectual content of an item, provided by the cataloguer
<support>	contains a description of the materials etc. which make up the physical support for the written part of a manuscript.
<supportDesc>	groups elements describing the physical support for the written part of a manuscript.
<surrogates>	contains information about any digital or photographic representations of the manuscript being described which may exist in the holding institution or elsewhere.
<textLang>	describes the languages and writing systems used by a manuscript (as opposed to its description, which is described in the langUsage element)
<watermark>	contains a word or phrase describing a watermark or similar device.
<width>	contains a measurement measured along the axis perpendicular to the spine.

msdescription Elements

http://projects.oucs.ox.ac.uk/teiweb/Drafts/catalogue.xml?ID=Module_msdescription

Βασική δομή ενός TEI αρχείου

Το TEI βασίζεται στην XML και επομένων μπορεί να αξιοποιήσει τις τεχνικές αποθήκευσης, αναζήτησης και ανάκτησης, διαχείρισης και εμφάνισης κωδικοποιημένων δεδομένων που αναπτύσσονται για τις βάσεις δεδομένων.

Η διαδικασία διάκρισης των δομικών, μορφολογικών και σημασιολογικών χαρακτηριστικών του κειμένου γίνεται χρησιμοποιώντας XML ετικέτες (tags) οι οποίες «χαρακτηρίζουν» το κείμενο σύμφωνα με τους κανόνες και τις οδηγίες ενός προτύπου.

Για τη δημιουργία και επεξεργασία TEI εγγράφων χρησιμοποιούνται επεξεργαστές XML εγγράφων (XML editors), όπως για παράδειγμα το <oXygen/> στον δικτυακό τόπο του οποίου δημοσιεύονται οδηγίες για την επεξεργασία TEI εγγράφων και παρέχονται έγγραφα μορφοποίησης (stylesheets) και οδηγοί (document templates) για την ευκολότερη επεξεργασία των εγγράφων.

Κάθε ηλεκτρονικό κείμενο που ακολουθεί το πρότυπο κωδικοποίησης TEI αποτελείται από τα παρακάτω στοιχεία:
<!ELEMENT TEI (teiHeader, text)>

- <teiHeader>, που τεκμηριώνει όλα τα σχετικά με το έγγραφο μεταδεδομένα.
- <text>, που περιλαμβάνει την πλήρη μεταγραφή του κειμένου και την ανάλυση του συνόλου των δεδομένων στα επιμέρους τμήματα που το συγκροτούν.

Το στοιχείο teiHeader αποτελείται από τα παρακάτω στοιχεία (είναι η ρίζα τους):
<!ELEMENT teiHeader (fileDesc, encodingDesc, profileDesc, revisionDesc)>

- <fileDesc> υποχρεωτικό, παρέχει μια πλήρη βιβλιογραφική περιγραφή του κωδικοποιημένου κειμένου στοιχεία τίτλου, έκδοσης και προέλευσης του.
- <encodingDesc> προαιρετικό, περιγράφει τη σχέση του κωδικοποιημένου κειμένου με την πηγή ή τις πηγές του. Παρέχει πληροφορίες για το εάν και πώς το κείμενο κανονικοποιήθηκε κατά τη μεταγραφή του και τις μεθόδους στις οποίες βασίστηκε η κωδικοποίησή του.
- <profileDesc> προαιρετικό, παρέχει πληροφορίες για τον τρόπο δημιουργίας του κειμένου, τις γλώσσες που χρησιμοποιούνται σε αυτό και τη θεματική του περιγραφή και ταξινόμηση. Ιδιαίτερα χρήσιμο στοιχείο σε περιπτώσεις σωμάτων ή μεγάλων συλλογών κειμένων γιατί συνεισφέρει στην αναζήτηση και

- ανάκτηση συγκεκριμένων κειμένων.
- <revisionDesc> προαιρετικό, χρησιμοποείται για την καταγραφή του ιστορικού των αλλαγών έγιναν και γίνονται στο κωδικοποιημένο κείμενο κατά την επεξεργασία του.

Στην Ελλάδα η αξιοποίηση σχετικών προτύπων περιγραφής και ψηφιακών εργαλείων διαχείρισης κειμένων για τις Ανθρωπιστικές σπουδές βρίσκεται σε πρώιμο στάδιο. Η παρουσία κωδικοποιημένων ελληνικών κειμένων σε ψηφιακές βιβλιοθήκες του εξωτερικού (Perseus Digital Library και EpiDoc) καταδεικνύει τη χρησιμότητα του TEI για την επεξεργασία των ελληνικών κειμένων.