Text Encoding Initiative: επισκόπηση, προβλήματα και εφαρμογές

Text Encoding Initiative: review, problems and real world implementations

Λίνα Μπουντούρη
Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ) / ΕΙΕ, Βασ. Κωνσταντίνου 48, 11635 Αθήνα
bountouri@ekt.gr

Lina Bountouri
NDC / NHRF
bountouri@ekt.gr

Περίληψη

Τα τελευταία χρόνια έχουν γίνει αρκετές προσπάθειες στο επιστημονικό πεδίο της Πληροφορικής για τις Ανθρωπιστικές σπουδές (Humanities Computing), με σκοπό τη δημιουργία προτύπων περιγραφής και εργαλείων διαχείρισης των κειμένων ανθρωπιστικών σπουδών στο Διαδίκτυο. Η δυνατότητα ηλεκτρονικής επεξεργασίας και έκδοσης κειμένων αποτέλεσε βασική εφαρμογή των εργαλείων της πληροφορικής στις Ανθρωπιστικές σπουδές και αποτέλεσε σημαντικό παράγοντα για την ανάδειξη, τη διαχείριση και την επιστημονική επεξεργασία των έργων του γραπτού λόγου.

Για την ηλεκτρονική εμφάνιση και διαχείριση των κειμένων ανθρωπιστικών σπουδών, οι ερευνητές του κλάδου χρησιμοποίησαν σε μεγάλο βαθμό τις γλώσσες σήμανσης (HTML, SGML, XML). Η SGML και η XML, λόγω της δομής τους, αποτέλεσαν εξ' αρχής μια πιο αποτελεσματική βάση – σε σχέση με την HTML – για τη δημιουργία προτύπων και μεταδεδομένων που αφορούν τα ηλεκτρονικά κείμενα.

Το συγκεκριμένο άρθρο ασχολείται με το πρότυπο Text Encoding Initiative (TEI), το οποίο εμφανίστηκε το 1987. Το TEI βασίστηκε αρχικά στην SGML και στη συνέχεια και στην XML, με στόχο να καθιερωθεί ως ένα διεθνές πρότυπο το οποίο θα χρησιμοποιείται από βιβλιοθήκες, μουσεία, εκδότες και ακαδημαϊκούς για την παρουσίαση λογοτεχνικών και γλωσσολογικών κειμένων σε ηλεκτρονική μορφή. Μέσα από τη γενική περιγραφή του TEI Document Type Definition (TEI DTD) και των συστατικών μερών του, θα εξακριβωθεί η χρησιμότητά του για τέτοιου είδους υλικό, ενώ θα δοθεί ιδιαίτερη έμφαση στη δομή και στη χρησιμότητα του TEI Header. Το TEI Header αποτελεί αναπόσπαστο κομμάτι του TEI DTD, καθώς παρέχει χρήσιμες πληροφορίες τόσο στους χρήστες που αναζητούν και ανακτούν ηλεκτρονικά κείμενα όσο και στους βιβλιοθηκονόμους ή στους συντάκτες των κειμένων αυτών.

Το TEI αξιοποιείται από φορείς πληροφόρησης του εξωτερικού, κυρίως από βιβλιοθήκες και αρχεία. Το γεγονός αυτό δίνει το έναυσμα για τη συνεχή έρευνα στο πεδίο δημιουργίας εργαλείων επεξεργασίας, εμφάνισης και αναζήτησης – ανάκτησης των TEI εγγράφων. Παράλληλα, η υιοθέτησή του σε πραγματικές εφαρμογές κάνει εμφανή τα πλεονεκτήματα της

χρήσης του και συγχρόνως αναδεικνύει τα ανοικτά ερευνητικά ερωτήματα για την περαιτέρω βελτίωση του. Οι πλευρές αυτές θα εξεταστούν στα πλαίσια του παρόντος άρθρου.

Συνοψίζοντας, η αναφορά σε παραδείγματα κωδικοποίησης ελληνικών κειμένων καταδεικνύει τόσο τη χρησιμότητα του TEI για τους Έλληνες ερευνητές των ανθρωπιστικών σπουδών όσο και το πρόσφορο έδαφος που υπάρχει στην ελληνική πραγματικότητα για την αξιοποίηση του προτύπου.

Λέξεις-κλειδιά: *Text Encoding Initiative, περιγραφικά μεταδεδομένα, Πληροφορική - Ανθρωπιστικές σπουδές.*

Abstract

The last decades, there have been many efforts in the scientific field of Humanities Computing in order to create standards and useful tools to edit, manage and deliver the humanities textual material in the web. Editing and delivering electronically the textual material was the basic implementation of Computing in the Humanities field. Those procedures were of major importance in promoting, managing and editing texts.

The humanists, in order to manage and represent electronically the humanities texts, have used the markup languages (HTML, SGML, and XML). The languages SGML and XML proved to be more efficient, because of their structure, for the creation of standards and metadata concerning the electronic textual material, in comparison to HTML.

The particular article examines the Text Encoding Initiative standard (TEI), which was established in 1987. TEI was initially based on SGML but, nowadays, there is also a TEI - XML compatible version. Its main target was to function as a community – based standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching. By giving a general description of the TEI Document Type Definition (TEI DTD) and its components, we will prove TEI's usefulness for humanities textual material. At the same time, we will emphasize on the use and importance of TEI Header, which is mandatory part of the TEI DTD. The TEI Header offers valuable information to the users that search and retrieve the electronic documents and, at the same time, to the librarians or the creators of the documents.

The TEI is used in a world wide basis, especially from libraries and archival institutions. This fact has led to a continuous research for the creation of tools to edit, represent, search and retrieve TEI documents. In parallel, the adoption of TEI in real world implementations makes obvious its advantages and, at the same time, its open research problems that have to be reconsidered in order to improve the standard. Those problematic areas will be analysed in the particular article.

Last but not least, the presentation of textual material encoded in TEI shows us not only the usefulness of TEI for the Greek humanists but also opportunity provided to exploit this standard in Greek humanities material.

Keywords: *Text Encoding Initiative, descriptive metadata, Humanities Computing.*

Πηγές

TEI Consortium (2003). "Text Encoding Initiative". [Web Page]. [Accessible to: http://www.tei-c.org]

Bountouri, Lina (2003). "Encoding the poems of Dionysios Solomos in TEI P4 (Dissertation submitted in partial fulfilment of the requirements of the degree of MSc of University College of London)".

1. Εισαγωγή

Την τελευταία δεκαετία, στα πλαίσια της εξέλιξης της Πληροφορικής για τις Ανθρωπιστικές σπουδές (Humanities Computing), έχουν παρατηρηθεί σημαντικές προσπάθειες για τη δημιουργία προτύπων κωδικοποίησης δεδομένων και εργαλείων διαχείρισης των κειμένων των Ανθρωπιστικών σπουδών. Οι προσπάθειες αυτές στοχεύουν στην εξυπηρέτηση των πληροφοριακών αναγκών των ερευνητών του σχετικού επιστημονικού πεδίου, δημιουργώντας αξιοποιήσιμες ηλεκτρονικές συλλογές. Η συγκεκριμένη ομάδα χρηστών έχει αυξημένες και εξειδικευμένες ανάγκες πρόσβασης, αναζήτησης και ανάκτησης τεκμηρίων, καθώς το ενδιαφέρον των χρηστών εστιάζεται κυρίως σε βιογραφίες και γραμματικές, συντακτικές, πραγματολογικές και ετυμολογικές σημειώσεις.

Ένα από τα πλέον αποτελεσματικά εργαλεία για την επεξεργασία και διαχείριση των συλλογών ανθρωπιστικού περιεχομένου είναι το *ηλεκτρονικό κείμενο* (electronic text). Ο όρος «ηλεκτρονικό κείμενο» χρησιμοποιείται για να δηλώσει κυρίως την κωδικοποίηση ενός κειμένου παρά την ηλεκτρονική ή ψηφιακή απεικόνισή του (Hockey, 2000, p.1).

Για τη δημιουργία, τη δομημένη εμφάνιση και την επεξεργασία των ηλεκτρονικών κειμένων αξιοποιούνται οι *γλώσσες σήμανσης* (markup languages). Η πρώτη γλώσσα σήμανσης που χρησιμοποιήθηκε για τη δημιουργία ηλεκτρονικών κειμένων είναι η Πρότυπη Γενικευμένη Γλώσσα Σήμανσης (Standard Generalized Markup Language - SGML).[1] Η SGML είναι μία γλώσσα που ορίζει μεθόδους ανεξάρτητες από υλικό και λογισμικό για την παρουσίαση κειμένων σε ηλεκτρονική μορφή, κωδικοποιώντας το περιεχόμενο και τη δομή τους. Από την SGML προέκυψε η γλώσσα της οποίας η χρήση επικρατεί στο Διαδίκτυο, η HyperText Markup Language (HTML).[2] Η HTML, αν και εύκολη στη χρήση, αναπαριστά την εμφάνιση του κειμένου και όχι το περιεχόμενο της πληροφορίας, γεγονός που περιορίζει τις δυνατότητές της. Η πολυπλοκότητα εκμάθησης και χρήσης της SGML, καθώς και η επιφανειακότητα της HTML, οδήγησαν στη δημιουργία της Επεκτάσιμης Γλώσσας Σήμανσης (eXtensible Markup Language - XML),[3] η οποία αποτελεί μία απλουστευμένη μορφή της SGML. Η SGML και η XML χαρακτηρίζονται ως «meta – markup languages», μπορούν δηλαδή να χρησιμοποιηθούν για τον ορισμό νέων λεξιλογίων και γλωσσών σήμανσης. Με βάση τη συγκεκριμένη τους ιδιότητα, καθώς και τη δομημένη σύνταξη που παρέχουν, η SGML και η XML αποτέλεσαν ισχυρή βάση για τη δημιουργία πρότυπων κωδικοποίησης δεδομένων σε ποικίλα επιστημονικά πεδία, όπως αυτό των Ανθρωπιστικών σπουδών.

[1] Διαθέσιμο στο: http://www.w3.org/MarkUp/SGML/
[2] Διαθέσιμο στο: http://www.w3.org/MarkUp/
[3] Διαθέσιμο στο: http://www.w3.org/XML/

2. Text Encoding Initiative (TEI)

Το μεγαλύτερο μέρος των πηγών που σχετίζεται με τις Ανθρωπιστικές σπουδές έχει τη μορφή κειμένου, γεγονός που εντείνει την ανάγκη δημιουργίας προτύπων και εργαλείων διαχείρισης για το συγκεκριμένο είδος. Τα κείμενα των ανθρωπιστικών σπουδών είναι πολύπλοκα και περιλαμβάνουν ποικιλία δεδομένων, τα οποία πρέπει να κωδικοποιηθούν με τον καλύτερο δυνατό τρόπο ώστε να βελτιωθεί η αναζήτηση και η ανάκτηση της πληροφορίας. Η πολυπλοκότητά τους έγκειται σε μεγάλο βαθμό στο πλήθος των διαφορετικών πληροφοριών που περιλαμβάνουν καθώς και στις μορφές κειμένων που μπορεί να συνδυάζουν (π.χ. ποίηση σε συνδυασμό με πεζό λόγο).

Το διεθνές πρότυπο Text Encoding Initiative (TEI)[4] ήλθε να αποτελέσει την κατάλληλη λύση για την κωδικοποίηση και ανταλλαγή κειμένων των Ανθρωπιστικών σπουδών. Το TEI εκδίδεται και υποστηρίζεται από το TEI Consortium, το οποίο παράγει σύνολα κανόνων και οδηγίες για την επεξεργασία και ανταλλαγή όχι μόνο κειμένων που αφορούν τις Ανθρωπιστικές σπουδές (π.χ. θεατρικά έργα και ποίηση) αλλά και όσων αφορούν την ευρύτερη γλωσσολογική βιομηχανία (π.χ. έντυπα λεξικά). Οι εκδόσεις του προτύπου εκφράζονται σε SGML, ενώ η πιο πρόσφατη (TEI P4) είναι συμβατή με την XML.

Το TEI ανήκει στα πρότυπα κωδικοποίησης δεδομένων που έχουν χρησιμοποιηθεί ιδιαίτερα στο περιβάλλον των βιβλιοθηκών και των αρχείων. Αποτελεί κατά κύριο λόγο περιγραφικό πρότυπο κωδικοποίησης δεδομένων, αλλά χαρακτηρίζεται και ως διαχειριστικό και δομικό πρότυπο (descriptive, administrative and structural metadata). (DLC/BAER Metadata group 3, 2002) Περιγραφικό πρότυπο, διότι περιγράφει πηγές με σκοπό την ανάκτηση και την ταυτοποίηση της πληροφορίας και διαχειριστικό, διότι παρέχει πληροφορίες για τη διαχείριση μίας πηγής, όπως πότε και από ποιον έχει δημιουργηθεί, τεχνικές λεπτομέρειες, πληροφορίες πρόσβασης κτλ. Τέλος, το TEI ανήκει και στα πρότυπα κωδικοποίησης δομικών μεταδεδομένων εφόσον υποδεικνύει τον τρόπο συνδυασμού των αντικειμένων ενός τεκμηρίου, π.χ. τη διάταξη των σελίδων προκειμένου να διαμορφώσουν κεφάλαια. (Hodge, 2001, p.3)

2.1. TEI DTD (P4)

Η τελευταία έκδοση του TEI (XML-compatible edition, P4) εκφράζεται, όπως και οι προηγούμενες εκδόσεις μέσα από ένα Document Type Definition, το οποίο ορίζει περίπου 450 ετικέτες για την κωδικοποίηση ηλεκτρονικών κειμένων. Το TEI P4 DTD είναι ένα από τα πιο πολύπλοκα και εξειδικευμένα DTD. Διαθέτει μία ποικιλία στοιχείων και γνωρισμάτων επιτρέποντας την αναλυτική κωδικοποίηση ενός κειμένου. Οι υποχρεωτικοί κόμβοι είναι περιορισμένοι σε αριθμό, γεγονός που συντείνει στην ευελιξία επιλογής και χρήσης κόμβων.

Αναλυτικά, το TEI P4 DTD περιλαμβάνει:

- Τις ετικέτες «core tag sets», οι οποίες αποτελούν βασικό μέρος του DTD και η χρήση τους είναι υποχρεωτική. Οι συγκεκριμένες ετικέτες προσδιορίζουν το ηλεκτρονικό κείμενο, δίνοντας πληροφορίες για τους δημιουργούς, τις πηγές, ενώ παράλληλα ορίζουν φαινόμενα που εμφανίζονται συχνά σε ένα κείμενο όπως παράγραφοι, ονόματα, διορθώσεις του επιμελητή, σημειώσεις κτλ.” (Sperberg-McQueen, 1994). Οι core tag sets ετικέτες ορίζονται μέσα από δύο DTD:

[4] Διαθέσιμο στο: http://www.tei-c.org/

- Το TEI.core.dtd. Οι ετικέτες που ορίζει αυτό το σύνολο μπορούν να εμφανιστούν σε οποιοδήποτε σημείο της δομής του κειμένου, όπως η παράγραφος.
- Το TEI.header.dtd. Κάθε TEI έγγραφο πρέπει να περιλαμβάνει το συγκεκριμένο σύνολο ετικετών, διότι αυτό αποτελεί την τεκμηρίωσή του, τα μεταδεδομένα του.

◊ Τις ετικέτες «base tag sets», οι οποίες αποτελούν τα βασικά δομικά μέρη για συγκεκριμένους τύπους κειμένων. (Erjavec, 2002). Ως αποτέλεσμα, ανάλογα με τη φύση του κειμένου (έντυπα λεξικά, ποίηση κτλ), επιλέγεται το κατάλληλο σύνολο ετικετών. Τα «base tag sets» χωρίζονται στις εξής κατηγορίες:

- TEI.prose, (πεζός λόγος)
- TEI.verse, (ποίηση)
- TEI.drama, (θεατρικά έργα)
- TEI.spoken, (προφορικά κείμενα)
- TEI.dictionaries, (έντυπα λεξικά)
- TEI.terminology, (ορολογία)

Στην περίπτωση που απαιτείται συνδυασμός των παραπάνω ετικετών λόγω του περιεχομένου και της δομής ενός κειμένου, χρησιμοποιούνται δύο επιπλέον «base tag sets»: α) το TEI.general, το οποίο επιτρέπει τη χρήση διαφορετικών «base tag sets» σε διάφορα τμήματα ενός κειμένου, εξασφαλίζοντας ότι κάθε τμήμα χρησιμοποιεί μόνο ένα «base tag set», και β) το TEI.mixed, το οποίο επιτρέπει την ανάμειξη στοιχείων από κάθε «base tag set» στα τμήματα ενός κειμένου (Sperberg-McQueen και Burnard, 2002a).

◊ Τις ετικέτες «additional tag sets», οι οποίες εξυπηρετούν διάφορους σκοπούς, όπως το TEI.linking, το οποίο παρέχει ετικέτες για τη δημιουργία συνδέσμων και το TEI.names.dates, το οποίο παρέχει ετικέτες για ονόματα και ημερομηνίες.

◊ Τις ετικέτες «user defined tags», οι οποίες χρησιμεύουν στη δημιουργία, μετατροπή και επέκταση των ετικετών σύμφωνα με τη φύση και τις ανάγκες κωδικοποίησης του υλικού.

Το TEI Consortium λαμβάνοντας υπόψη ότι το TEI P4 DTD είναι περίπλοκο και περιλαμβάνει ετικέτες που χρησιμοποιούνται σπάνια στην κωδικοποίηση ενός κειμένου, συστήνει μία απλοποιημένη εκδοχή του, το TEI Lite DTD.[5] Το TEI Lite περιλαμβάνεται σε ένα αρχείο και έχει ήδη αξιοποιηθεί από μεγάλο αριθμό φορέων και προγραμμάτων, για παράδειγμα, από το Oxford Text Archive.[6]

2.2. TEI Header

Κάθε ηλεκτρονικό κείμενο κωδικοποιημένο σε TEI περιλαμβάνει, όπως προαναφέρθηκε, ένα σύνολο ετικετών που ονομάζεται TEI Header. Σκοπός του TEI Header είναι να κωδικοποιήσει βιβλιογραφικές και περιγραφικές πληροφορίες που αφορούν το ηλεκτρονικό έγγραφο.

Το TEI Header χαρακτηρίζεται ως «ηλεκτρονική σελίδα τίτλου» και αποτελείται από τέσσερα βασικά μέρη:

[5] Διαθέσιμο στο: http://www.tei-c.org/Lite/
[6] Διαθέσιμο στο: http://ota.ahds.ac.uk/

- ◊ Το στοιχείο *File Description* (<fileDesc>), το οποίο περιλαμβάνει πληροφορίες που λειτουργούν ως περιγραφικά μεταδεδομένα του TEI εγγράφου. Το συγκεκριμένο στοιχείο είναι υποχρεωτικό και περιγράφει το τεκμήριο με σκοπό να διευκολύνει τόσο την ταυτοποίησή του όσο και την αναζήτηση – ανάκτηση.
- ◊ Το στοιχείο *Encoding Description* (<encodingDesc>), το οποίο περιέχει πληροφορίες για την κανονικοποίηση του κειμένου, τα προβλήματα που προέκυψαν κατά την κωδικοποίηση, το επίπεδο κωδικοποίησης και ανάλυσης που εφαρμόζεται κτλ. (Seaman, 2000).
- ◊ Το στοιχείο *Profile Description* (<profileDesc>), το οποίο περιλαμβάνει μη βιβλιογραφικά δεδομένα, όπως πληροφορίες για τις γλώσσες που χρησιμοποιούνται στο κείμενο.
- ◊ Το στοιχείο *Revision Description* (<revisionDesc>), το οποίο επιτρέπει στους δημιουργούς των TEI εγγράφων να καταγράφουν το ιστορικό των αλλαγών που πραγματοποιούνται κατά την επεξεργασία του ηλεκτρονικού εγγράφου. (Sperberg-McQueen και Burnard, 2002b, p.79)

Το TEI Header αποτελεί αναπόσπαστο και ιδιαίτερα χρήσιμο μέρος του TEI εγγράφου. Η χρησιμότητά του έγκειται στις πληροφορίες που παρέχει στους χρήστες, στους ίδιους τους δημιουργούς των εγγράφων και στους βιβλιοθηκονόμους. Ενδεικτικά αναφέρεται ότι έχει χρησιμοποιηθεί ως βάση για τη δημιουργία αντίστοιχων συνόλων κόμβων, όπως του στοιχείου <eadheader> του Encoded Archival Description.[7]

Παράλληλα, το TEI Header παρέχει βιβλιογραφικές πληροφορίες οι οποίες αντιστοιχούν σε πεδία των διάφορων υλοποιήσεων του ISO 2709 και σε περιοχές των Άγγλο – Αμερικάνικων Κανόνων Καταλογογράφησης (AACR2). Για παράδειγμα:

TEI ετικέτες	**AACR2 περιοχές**	**MARC πεδία**
<titleStmt> <title>	9.1B-E 9.7B4	245 a 246
<projectDesc> <p>	9.7B6	500

Το TEI Header μοιράζεται μία κοινή βάση με τα βιβλιοθηκονομικά πρότυπα, γεγονός που διευκολύνει το πέρασμα από το ένα πρότυπο στο άλλο συντελώντας στην επικοινωνία και τη διαλειτουργικότητα ανάμεσα στα πρότυπα κωδικοποίησης δεδομένων.

Επιπλέον, από το TEI Header είναι εφικτή η αυτόματη δημιουργία μίας βιβλιογραφικής εγγραφής. Μία σχετική προσπάθεια πραγματοποιείται από το Oxford Text Archive, στα πλαίσια του οποίου μετατρέπονται αυτόματα τα TEI Header σε MARC εγγραφές, οι οποίες στη συνέχεια καταχωρούνται στον OPAC του Πανεπιστημίου της Οξφόρδης (Morrison, 1999).

Το TEI Header, δίνοντας βασικές βιβλιογραφικές ή μη πληροφορίες χρησιμοποιείται για την ευκολότερη διαχείριση των συλλογών. Μέσα από αυτό διευκολύνονται οι διαδικασίες αναζήτησης, επιλογής και πρόσβασης σε ένα ηλεκτρονικό κείμενο. Αναλυτικότερα, το

[7] Διαθέσιμο στο: http://www.loc.gov/ead/

στοιχείο File Description περιλαμβάνει στοιχεία, όπως το <title> και το <author>, τα οποία αποτελούν βασική πηγή ανάκτησης πληροφοριών. Επιπρόσθετα, υπάρχουν υπό – στοιχεία του File Description όπως το <sourceDesc> και το <editionStmt> που συνδέουν το ηλεκτρονικό έγγραφο με την έντυπη μορφή του δίνοντας εκτεταμένες πληροφορίες για αυτήν, καθώς και υπό – στοιχεία, όπως το <notesStmt>, που περιλαμβάνουν πληροφορίες για το περιεχόμενο ή τη φυσική μορφή του τεκμηρίου. Ένα ακόμη βασικό χαρακτηριστικό του TEI Header είναι ότι παρέχει πληροφορίες πρόσκτησης και πρόσβασης στο περιγραφόμενο υλικό, για παράδειγμα πληροφορίες δανεισμού, διαθεσιμότητας και απομακρυσμένης πρόσβασης (Shieh, 1998). Συγχρόνως, μπορεί να αξιοποιηθεί για τη δημιουργία εξειδικευμένων σημείων πρόσβασης, όπως η γλώσσα, η χρονική περίοδος, ακόμη και ο τύπος ενός αρχείου (Morrison, 1998).

3. Εφαρμογές και εργαλεία

Τα εργαλεία που σχετίζονται με τη δημιουργία και επεξεργασία των TEI εγγράφων μπορούν να χωριστούν σε α) εργαλεία για τη δημιουργία TEI DTD και TEI XML Schema, και σε β) επεξεργαστές XML εγγράφων (XML editors).

Το πρώτο εργαλείο για τη δημιουργία TEI DTD ονομάζεται "Pizza Chef"[8] υποστηρίζεται από το TEI Consortium και υλοποιεί την P4 έκδοση του προτύπου. Ωστόσο, η πολυπλοκότητα στη χρήση του οδήγησε στην αντικατάστασή του από το "Roma". Το "Roma" είναι ένα νέο εργαλείο για τη δημιουργία TEI κανόνων σε μορφή DTD, Relax NG ή W3C Schema. Αν και σε πρώιμη μορφή, το "Roma" έχει ήδη ξεπεράσει σε μεγάλο βαθμό τα προβλήματα του "Pizza Chef", αυτοματοποιώντας διαδικασίες και δημιουργώντας σύνολα κανόνων για τη δημιουργία TEI εγγράφων. Βρίσκεται σε πειραματικό στάδιο και υποστηρίζει την P4 έκδοση του προτύπου καθώς και την υπό επεξεργασία P5 έκδοσή του.

Για τη δημιουργία TEI εγγράφων χρησιμοποιούνται εργαλεία συγγραφής SGML και XML, καθώς και επεξεργαστές κειμένου (π.χ. UltraEdit). Οι επεξεργαστές που χρησιμοποιούνται κατά κύριο λόγο, σύμφωνα με πρόσφατη έρευνα του TEI Consortium, είναι οι: α)XMetaL,[9] β)Emacs,[10] και γ)<oXygen/>[11] (Thijs van den Broek, 2004). Αξίζει να σημειωθεί ότι στο δικτυακό τόπο του επεξεργαστή <oXygen/> δημοσιεύονται οδηγίες για την επεξεργασία TEI εγγράφων μέσω <oXygen/>. Επιπλέον, ο συγκεκριμένος επεξεργαστής παρέχει έγγραφα μορφοποίησης (stylesheets) και οδηγούς (document templates) για την ευκολότερη επεξεργασία των εγγράφων (Oxygen, 2003).

Το TEI βασιζόμενο στην SGML και στην XML μπορεί να αξιοποιήσει τις τεχνικές αποθήκευσης, αναζήτησης και ανάκτησης που αναπτύσσονται για τις γλώσσες σήμανσης, όπως τις SGML και XML βάσεις δεδομένων. Παράλληλα, χρησιμοποιώντας ως βάση τις SGML και XML τεχνικές που προαναφέρθηκαν, αναπτύσσονται μέσα αποθήκευσης και διαχείρισης για δεδομένα κωδικοποιημένα σε TEI όπως το teiPublisher, το οποίο χρησιμεύει στην αποθήκευση, αναζήτηση και εμφάνιση κειμένων κωδικοποιημένων σε TEILite. (Kumar κ.ά., 2004)

[8] Διαθέσιμο στο: http://www.tei-c.org/pizza.html
[9] Διαθέσιμο στο: http://www.tei-c.org/Software/Survey_XML_Software.pdf
[10] Διαθέσιμο στο: http://www.emacs.org/
[11] Διαθέσιμο στο: http://www.oxygenxml.com/

4. Πλεονεκτήματα και Προβληματικές περιοχές

Το Text Encoding Initiative, ως υλοποίηση της SGML και της XML, κληρονομεί τις ιδιότητες των δύο γλωσσών σήμανσης ενώ παράλληλα αξιοποιεί τις υποστηρικτικές τεχνολογίες και τα εργαλεία που τις περικλείουν. Το TEI είναι ένα πρότυπο ανεξάρτητο από πλατφόρμες, συστήματα και εφαρμογές ενώ συμβαδίζει με όλα τα πρωτόκολλα δικτύου για την ανταλλαγή της πληροφορίας.

Ακολουθώντας τις σύγχρονες τάσεις που επηρεάζουν τη μορφή των προτύπων κωδικοποίησης δεδομένων, το TEI περνά πλέον σε μορφή XML. Η χρήση μίας γλώσσας σήμανσης, όπως είναι η XML, για την έκφραση των προτύπων έχει ιδιαίτερα θετικές συνέπειες στη διαχείριση της ηλεκτρονικής πληροφορίας. Η ανεξάρτητη φύση της XML επιτρέπει την αξιοποίηση ήδη υπαρχόντων τεχνολογιών και εργαλείων (π.χ. επεξεργαστές κειμένου) για την επεξεργασία ηλεκτρονικών κειμένων.

Αξιοποιώντας τις υποστηρικτικές τεχνολογίες της XML, ένα έγγραφο TEI μπορεί να παρουσιαστεί σε ποικίλες μορφές με τη βοήθεια της eXtensible Stylesheet Language[12], γεγονός που είναι ιδιαίτερα χρήσιμο για τους ερευνητές των Ανθρωπιστικών Σπουδών. Η XSL επιτρέπει τη μορφοποίηση ενός TEI εγγράφου με ή χωρίς σημειώσεις, με ή χωρίς τις διαγραμμένες λέξεις από το συγγραφέα, σε μορφή εικόνας ή μορφή κειμένου κτλ.

Βασικό πλεονέκτημα του TEI P4 DTD είναι ότι επιδέχεται παραμετροποίηση και επέκταση, με σκοπό να ανταποκριθεί σε ιδιαίτερες ανάγκες περιγραφής τεκμηρίων (Sperberg-McQueen και Burnard, 2002c).

Τέλος, το TEI είναι ένα πρότυπο που έχει αυξημένες δυνατότητες στο να αποτυπώσει διαφόρων ειδών πληροφορίες ενός κειμένου, όπως σημειώσεις περιθωρίου. Επιτρέπει παράλληλα τη χρήση τυπογραφικών και σημασιολογικών κόμβων. Για παράδειγμα, το ζευγάρι γνωρίσματος – τιμής rend="italics" χρησιμοποιείται για να δηλώσει ότι το κείμενο που περιλαμβάνει είναι γραμμένο ή πρέπει να εμφανιστεί με «italics».

Η χρήση του TEI αποδεικνύεται αποδοτική ως προς την κωδικοποίηση κειμένων της ευρύτερης γλωσσολογικής βιομηχανίας. Παρ' όλα αυτά, η αξιοποίησή του σε πραγματικές εφαρμογές κάνει εμφανή τα αδύνατα σημεία του τα οποία αναλύονται στη συνέχεια.

4.1. Επικαλυπτόμενες ιεραρχίες

Ένα βασικό θεωρητικό πρόβλημα, το οποίο πηγάζει από τη δομή της SGML και της XML, είναι οι *επικαλυπτόμενες ιεραρχίες* (overlapping hierarchies). Η SGML και η XML έκδοση του Text Encoding Initiative κωδικοποιούν τα κείμενα ως ένα σύνολο ιεραρχιών, όπου τα στοιχεία περιλαμβάνουν υπό – στοιχεία δίχως ποτέ να επικαλύπτονται, υιοθετώντας την άποψη του DeRose (1990) ότι το κείμενο είναι μία ορισμένη ιεραρχία από αντικείμενα. Με άλλα λόγια, ένα βιβλίο αποτελείται από κεφάλαια, ένα κεφάλαιο από υπό – κεφάλαια κτλ. Η συγκεκριμένη διατύπωση είναι λογική, ωστόσο η πρακτική εφαρμογή του TEI καταδεικνύει την ύπαρξη πολλαπλών ιεραρχικών δομών με στοιχεία που επικαλύπτονται.

Το φαινόμενο των επικαλυπτόμενων ιεραρχιών (overlapping hierarchies) προκύπτει από τη συνύπαρξη διαφορετικών δομών σε ένα κείμενο, όπως της λογικής με τη φυσική δομή ή της δομής του κειμένου με τη φυσική δομή. Για παράδειγμα, ένα γραμματικό φαινόμενο, ένα απόσπασμα ή μία αναφορά μπορεί να ξεκινά σε μία γραμμή και να τελειώνει σε μία άλλη.

[12] Διαθέσιμο στο: http://www.w3.org/Style/XSL/

<l n="1"> Είπες<quotation>«Θα πάγω σ' άλλη γη, θα πάγω σ' άλλη θάλασσα</l>

<l n="2">...</l>

…

<l n="8">που τόσα χρόνια πέρασα και ρήμαξα και χάλασα.»</quotation></l>

Οι επικαλυπτόμενες ιεραρχίες δεν επιτρέπουν την εύκολη διαχείριση της πληροφορίας σε επίπεδο αποθήκευσης, αναζήτησης – ανάκτησης και μορφοποίησης των δεδομένων, λόγω της δυσκολίας στον διαχωρισμό τους. Παράλληλα, οι επικαλυπτόμενες ιεραρχίες δυσκολεύουν τους χρήστες που κωδικοποιούν τα κείμενα στο διαχωρισμό των λογικών και των δομικών μονάδων του τεκμηρίου και, κατά συνέπεια, στην απόδοση των κατάλληλων κόμβων. Στο γεγονός αυτό συντείνει η διαφορετική αντίληψη των ατόμων που ασχολούνται με την κωδικοποίηση για το κείμενο που επεξεργάζονται αλλά συχνά και η ίδια η δομή του κειμένου. Για παράδειγμα, στα νεότερα χειρόγραφα είναι δύσκολο να οριστεί η μονάδα γραφής, αφού οι συγγραφείς διακόπτουν απότομα το γράψιμο, αφήνουν προτάσεις ημιτελείς κ.τ.λ. (Vanhoutte, 2002).

Το TEI Consortium έχει προτείνει ένα αριθμό λύσεων για την αντιμετώπιση του συγκεκριμένου προβλήματος, όπως τη χρήση κενών στοιχείων (milestones). Τα κενά στοιχεία, όπως το <qb> και το <qe> για την αρχή και το τέλος μίας αναφοράς, δεν περιλαμβάνουν κείμενο όμως ορίζουν όρια γύρω από αυτό. Επιπρόσθετα, προτείνεται ο «τεμαχισμός» (Fragmentation) ενός στοιχείου σε μικρότερα προκειμένου να αποφεύγονται οι επικαλυπτόμενες ιεραρχίες. Τέλος, υπάρχουν τα «εικονικά στοιχεία» (virtual elements), τα οποία ενώνουν κείμενο που δεν μπορεί να δομηθεί ιεραρχικά ως ένα σύνολο. Εν τούτοις, σε κάθε λύση που προτείνεται «υποτιμάται» η δομή που συμπεριλαμβάνεται στα πλαίσια αυτής της λύσης. Το TEI Consortium με σκοπό να αντιμετωπίσει το πρόβλημα και να χαράξει μία συγκεκριμένη πολιτική έχει δημιουργήσει λίστα συζητήσεων για τις επικαλυπτόμενες ιεραρχίες (TEI Overlapping Markup SIG discussion list). [13]

4.2. Κανόνες περιγραφής – TEI Header

Η παρουσία του TEI Header είναι, όπως προαναφέρθηκε, απαραίτητη σε ένα TEI έγγραφο, διότι παρέχει βασικές πληροφορίες και αποτελεί ένα δίαυλο επικοινωνίας με τα υπόλοιπα πρότυπα κωδικοποίησης δεδομένων. Ωστόσο, έχουν διατυπωθεί προβληματισμοί τόσο για τις πληροφορίες που περιλαμβάνει όσο και για τον τρόπο καταγραφής αυτών.

Αρχικά, η μη καταγραφή απαραίτητων δεδομένων στο TEI Header, για παράδειγμα εγγραφών καθιερωμένου τύπου ή πληροφοριών απαραιτήτων για συγκεκριμένους φορείς, έχει ως αποτέλεσμα τη δημιουργία ελλιπών ή ακόμη και λανθασμένων μεταδεδομένων για τα TEI έγγραφα.

Συγχρόνως, η έλλειψη κανόνων για την περιγραφή των δεδομένων αποτελεί το βασικότερο πρόβλημα για την επεξεργασία του TEI Header. Ποικίλες απόψεις έχουν διατυπωθεί για την επίλυση του προβλήματος, όπως η υιοθέτηση των AACR2 rev. κανόνων για το στοιχείο Source Description, το οποίο και περιλαμβάνει πληροφορίες κυρίως για το έντυπο υλικό, και των ISBD (ER) κανόνων για τα υπόλοιπα υπό – στοιχεία του File Description, στα οποία περιγράφεται το ηλεκτρονικό έγγραφο. Παράλληλα, ο συγκεκριμένος παράγοντας επηρεάζει το βαθμό αξιοπιστίας των μετατροπών από το TEI Header σε πλήρεις και έγκυρες

[13] Διαθέσιμο στο: http://listserv.brown.edu/tei-ol-sig.html

βιβλιογραφικές εγγραφές (UKOLN, 2004). Παρ' όλα αυτά, το TEI Consortium δεν έχει ορίσει κατευθυντήριες γραμμές και πολιτική για το ζήτημα.

4.3. TEI P5

Η νέα έκδοση του Text Encoding Initiative αναμένεται να λύσει πρακτικά και θεωρητικά προβλήματα. Η έκφραση του προτύπου σε Document Type Definition μορφή περιορίζει σε μεγάλο βαθμό τις δυνατότητες έκφρασης συγκεκριμένων τιμών για το περιεχόμενο των κόμβων και τη χρήση υποστηρικτικών τεχνολογιών και προτύπων της XML (π.χ. "XML Namespaces").[14] Η P5 νέα έκδοση του προτύπου αναμένεται να επιλύσει το παραπάνω πρόβλημα με την έκφραση του προτύπου σε RelaxNG XML Σχήμα.[15]

5. Η Πληροφορική για τις Ανθρωπιστικές σπουδές στην Ελλάδα

Στην Ελλάδα το πεδίο της Πληροφορικής για τις Ανθρωπιστικές σπουδές βρίσκεται σε πρώιμο στάδιο. Υπάρχει, ωστόσο, ακαδημαϊκή παράδοση και πλούσιο υλικό προς εκμετάλλευση στο χώρο των Ανθρωπιστικών Σπουδών, παράγοντες που συντείνουν στην αξιοποίηση σχετικών εργαλείων και προτύπων. Η παρουσία κωδικοποιημένων ελληνικών κειμένων σε ψηφιακές βιβλιοθήκες του εξωτερικού (Perseus Digital Library[16] και EpiDoc[17]) καταδεικνύει σε μεγάλο βαθμό τη χρησιμότητα του TEI για την επεξεργασία των ελληνικών κειμένων.

Ενδεικτικά αναφέρεται ένα παράδειγμα κωδικοποίησης ελληνικών κειμένων και η αντίστοιχη μορφοποιημένη προβολή του με τη χρήση ενός XSL εγγράφου.

Παράδειγμα 1

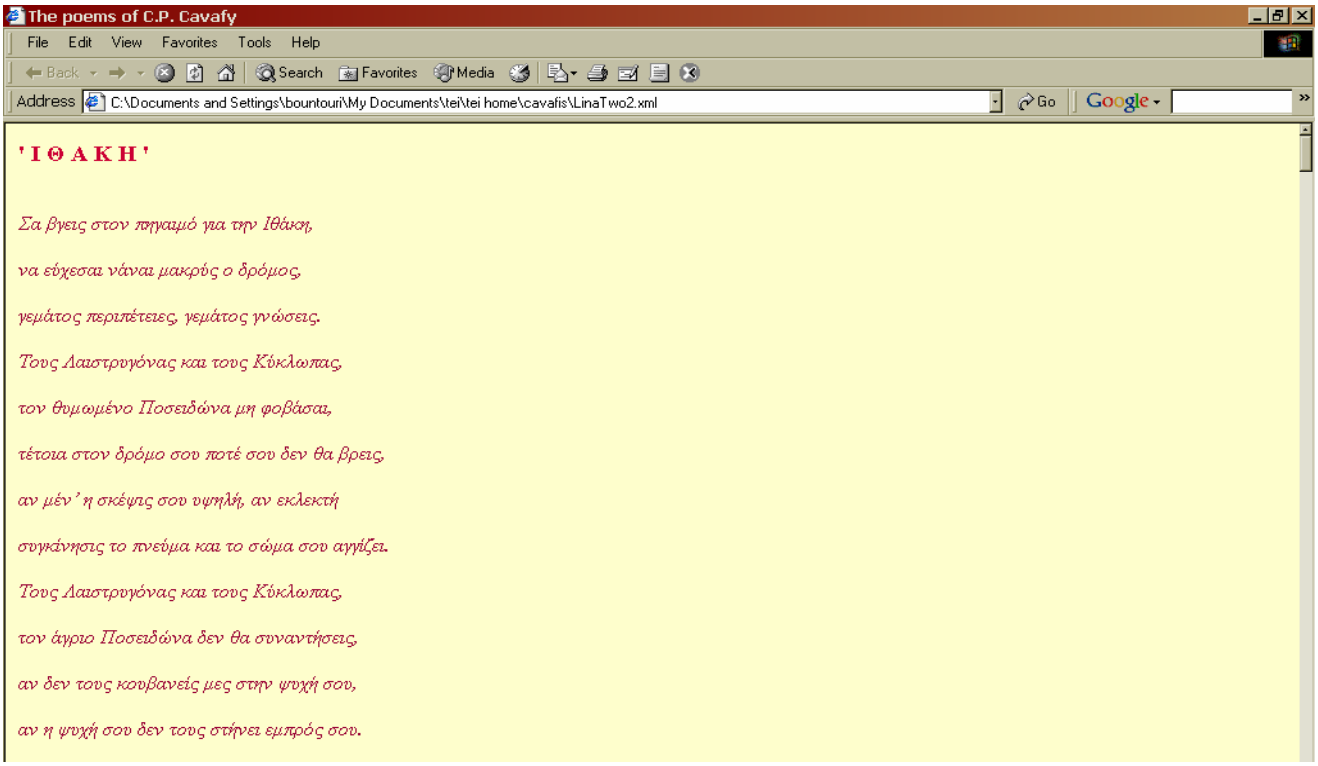
Κωδικοποιημένη μορφή

```
<TEI.2>
<teiHeader>
        <fileDesc>
                <titleStmt>
                        <title>Τα ποιήματα του Κ.Π. Καβάφη: ηλεκτρονική έκδοση</title>
                        <author> Κ.Π. Καβάφης</author>
                </titleStmt>
                <publicationStmt>
                        <publisher>Εκδόσεις Ίκαρος</publisher>
                        <date>1966</date>
                </publicationStmt>
        </fileDesc>
        <revisionDesc>
                <change n="1.00">
                        <date value="2004-15-10">15 Οκτωβρίου 2004</date>
                        <respStmt>
                                <name>Λίνα Μπουντούρη</name>
```

[14] Τα XML Namespaces είναι μία συλλογή ονομάτων, προσδιορισμένα από μία URI αναφορά, τα οποία χρησιμοποιούνται στα XML έγγραφα ως τύποι στοιχείων και ονόματα γνωρισμάτων. Αξιοποιούνται προκειμένου να διαχωρίσουν στοιχεία και γνωρίσματα προέρχονται από διαφορετικά λεξιλόγια και έχουν διαφορετική σημασιολογία ωστόσο μοιράζονται το ίδιο όνομα.
[15] Διαθέσιμο στο: http://www.tei-c.org/P5/
[16] Διαθέσιμο στο: http://www.perseus.tufts.edu/
[17] Διαθέσιμο στο: http://www.unc.edu/awmc/epidoc/

```
                                        <resp>encoder</resp>
                                </respStmt>
                                <item>Διορθώσεις ημιτελείς από το
                                        <date value="2004-03-05">Μάρτη 2004</date>
                                </item>
                        </change>
                </revisionDesc>
</teiHeader>
<text>
        <body>
                <lg type="free">
                        <l>Σα βγεις στον πηγαιμό για την Ιθάκη, </l>
                        <l> να εύχεσαι νάναι μακρύς ο δρόμος, </l>
                        <l> Τους Λαιστρυγόνας και τους Κύκλωπας, </l>
                        <l>τον θυμωμένο Ποσειδώνα μη φοβάσαι,</l>
                        <l>τέτοια στον δρόμο σου ποτέ σου δεν θα βρεις,</l>
                        <l>αν μέν' η σκέψις σου υψηλή, αν εκλεκτή</l>
                        <l>συγκίνησις το πνεύμα και το σώμα σου αγγίζει.</l>
                        <l>Τους Λαιστρυγόνας και τους Κύκλωπας,</l>
                        <l>τον άγριο Ποσειδώνα δεν θα συναντήσεις,</l>
                        <l>αν δεν τους κουβανείς μες στην ψυχή σου,</l>
                        <l>αν η ψυχή σου δεν τους στήνει εμπρός σου.</l>
                </lg>
        </body>
</text>
</TEI.2>
```

Προβολή μορφοποίησης



6. Επίλογος

Η αξιοποίηση του Text Encoding Initiative αποδεικνύεται ιδιαίτερα αποτελεσματική στην κωδικοποίηση κειμένων των Ανθρωπιστικών Σπουδών δημιουργώντας προσβάσιμες και αναζητήσιμες συλλογές. Το TEI ξεπέρασε κατά πολύ τον αρχικό του στόχο που ήταν η ανταλλαγή κειμένων και αποτελεί πλέον μέσο για τη μόνιμη αποθήκευση κειμένων καθώς και για την πρόσβαση στο υλικό. Ωστόσο, υπάρχουν ακόμη ανοιχτά ερευνητικά ζητήματα, όπως οι επικαλυπτόμενες ιεραρχίες και η έλλειψη κανόνων περιγραφής των δεδομένων, τα οποία απαιτούν περαιτέρω θεώρηση από το TEI Consortium και τη διεθνή ερευνητική κοινότητα.

Βιβλιογραφία

Broek, Thijs van den (2004) "Survey on XML editing software" [Διαθέσιμο στο http://www.tei-c.org/Software/Survey_XML_Software.pdf]

DeRose, S. J., κ.ά. (1990) "What is Text, Really?" *Journal of Computing in Higher Education,* 1 (2), pp. 3-26.

DLC/BAER Metadata group 3 (2002) "Metadata standards feature overview" [Διαθέσιμο στο http://staffweb.library.northwestern.edu/dl/metadata/standardsinventory/saimsgrid.xls]

Erjavec, Tomaz (2002) "TEI and GENIA" [Διαθέσιμο στο http://nl.ijs.si/et/talks/tei-genia/#intro]

Hockey, Susan (2000) *Electronic texts in the humanities: principles and practice.* Oxford University Press.

Hodge, Gail (2001). *Metadata: made simpler*. NISO.

Kumar, Amit κ.ά. (2004) "teiPublisher a repository management system for TEI documents" [Διαθέσιμο στο http://www.hum.gu.se/allcach2004/AP/html/prop118.html]

Morrison, Alan (1999) "Delivering Electronic Texts Over the Web: The Current and Planned Practices of the Oxford Text Archive" *Computers and the Humanities,* 33 (1) [Διαθέσιμο στο http://www.kluweronline.com/issn/0010-4817]

Morrison, Alan, Popham, Michael and Karen Wikander (1998) "Creating and documenting electronic texts: document analysis" [Διαθέσιμο στο http://ota.ahds.ac.uk/documents/creating/chap6.html]

Oxygen (2003) "How to create a TEI XML document and convert it to PDF within the <oXygen/> XML Editor" [Διαθέσιμο στο http://www.oxygenxml.com/doc/HowToCreatePDFUsingTEI.pdf]

Seaman, David (2000) "Guidelines for SGML Text Mark-up at the Electronic Text Center" [Διαθέσιμο στο http://etext.lib.virginia.edu/tei/uvatei4.html]

Shieh, Jackie (1998). "The TEI Header and the Cataloguing Rules" [Διαθέσιμο στο http://www.ala.org/Content/NavigationMenu/ALCTS/Division_groups/MARBI/Next_Section_3.htm#tei]

Sperberg-McQueen C.M. and Burnard, Lou (2002a) "Structure of the TEI Document Type Definition" [Διαθέσιμο στο http://www.tei-c.org/P4X/SG.html]

Sperberg-McQueen C.M. and Burnard, Lou (2002b) *Guidelines for Electronic Text Encoding and Interchange*. Vol.1. TEI Consortium.

Sperberg-McQueen C.M. and Burnard, Lou (2002c) “Modifying and Customizing the TEI DTD” [Διαθέσιμο στο http://www.tei-c.org/P4X/MD.html]

Sperberg-McQueen, C.M. (1994) “Textual criticism and the Text Encoding Initiative” [Διαθέσιμο στο http://www.tei-c.org/Vault/XX/mla94.html]

UKOLN (2004) “QA In The Construction Of A TEI Header” [Διαθέσιμο στο http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-69/]

Vanhoutte, Edward (2002) “Texts and transcriptions: mapping scribal complexities onto a line of text” [Διαθέσιμο στο http://www.kantl.be/ctb/vanhoutte/pub/2002/drh02abstr.htm]