Μετανάστευση τεκμηρίων από το πρότυπο TEI P3 (SGML) σε P4 (XML)

Migration of TEI documents from TEI P3 (SGML) to P4 (XML) standard

Γεωργία Αποστολοπούλου
Ιόνιο Πανεπιστήμιο (georgia@ionio.gr)

Ανέστης Σίτας
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (sitas@lit.auth.gr)

Περίληψη

Το πρότυπο TEI (Text Encoding Initiative) εμφανίστηκε σχετικά πρόσφατα και αξιοποιήθηκε ευρέως από βιβλιοθήκες και υπηρεσίες πληροφόρησης, με αποτέλεσμα σύντομα να αναχθεί σε διεθνές πρότυπο κωδικοποίησης των πηγών σε μορφή κειμένου. Οι τρεις πρώτες εκδόσεις του (TEI P1-3), που κυκλοφόρησαν από το 1990 έως το 1999, ήταν συμβατές με τη γλώσσα SGML. Η τέταρτη έκδοσή του κυκλοφόρησε με κύριο χαρακτηριστικό τη συμβατότητα με την XML. Τέλος, στις αρχές του 2005 κυκλοφόρησε η πιο πρόσφατη έκδοση (TEI P5). Η μετανάστευση των τεκμηρίων TEI από την έκδοση P3 στην P4 είναι απαραίτητη, καθώς η άμεση μετάβαση από την P3 στην P5 θα είναι εξαιρετικά δύσκολη. Το TEI, δεν σκοπεύει να προσφέρει μια τέτοια δυνατότητα, επομένως, η διαδικασία μετανάστευσης από την έκδοση P3 στην P4, θα συνεχιστεί για πολλά ακόμη χρόνια και κατά συνέπεια θα εξακολουθεί να χρησιμοποιείται η τέταρτη έκδοσή του. Στην εργασία αυτή θα ασχοληθούμε με την διαδικασία μετανάστευσης των τεκμηρίων της μορφής TEI και με τα ζητήματα που προκύπτουν από αυτή.

Abstract

The TEI (Text Encoding Initiative) standard was relatively recently developed and adopted widely by libraries and information services. Shortly, it became an international encoding standard for full text sources. TEI's first three editions (TEI P1-3), released between 1990 and 1999, were compatible with SGML language. It's 4th edition come up, with main characteristic the XML compatibility. Finally, in the beginning of 2005 the most recent edition, TEI P5, was released. A whole skeptical area in TEI, it concerns the 4th edition (P4) and it deals with the conversion of TEI documents from SGML P3 edition to XML P4. This kind of migration is essential. There will be no direct passage from P3 to P5. TEI does not intend to offer such a possibility, consequently, the process of migration from edition P3 to P4, will be continued for many years. That leads us to the fact, that the 4th edition of TEI, will continue to be used. In this paper we will examine the process of migration from TEI P3 documents to P4 and the upcoming questions.

Εισαγωγή

Η ευρεία χρήση της τεχνολογίας, των προσωπικών υπολογιστών και του Διαδικτύου κατά τη διάρκεια των δύο τελευταίων δεκαετιών, πέρα από την επιρροή και τις αλλαγές που επέφεραν στον τομέα των θετικών επιστημών, δεν μπορούσαν να αφήσουν ανεπηρέαστο και τον τομέα των ανθρωπιστικών επιστημών. Η χρήση και η εφαρμογή των εργαλείων και των μεθόδων της επιστήμης της πληροφορικής στον χώρο των ανθρωπιστικών επιστημών είχε ως αποτέλεσμα τη δημιουργία ενός νέου

γνωστικού πεδίου, της «**Πληροφορικής των Ανθρωπιστικών Επιστημών**» (Humanities Computing). Αντικείμενο της έρευνάς του είναι η υποστήριξη των νέων τεχνολογιών, με σκοπό τη δημιουργία νέων μεθόδων και εργαλείων για την αποτελεσματικότερη ικανοποίηση των πληροφοριακών αναγκών των ερευνητών του χώρου αυτού (Unsworth, 2002). Τα δημιουργούμενα εργαλεία δεν περιορίζονται μόνο στην ηλεκτρονική απεικόνιση και την ηλεκτρονική επεξεργασία, αλλά κυρίως στοχεύουν στην οργάνωση, τη διαχείριση και τη διάθεση των πηγών μέσω του Διαδικτύου.

Κωδικοποίηση

Είναι εμφανές πως όταν αναφερόμαστε σε πηγές ανθρωπιστικών, πολύ δε περισσότερο φιλολογικών σπουδών, αναφερόμαστε συνήθως σε πηγές που βρίσκονται σε μορφή κειμένου. Στα φιλολογικά κείμενα σημαίνουσα σημασία έχουν τα γραμματικά, συντακτικά, πραγματολογικά, ετυμολογικά και άλλα φαινόμενα. Συνεπώς οι προσεγγίσεις που αναπτύσσονται τείνουν να προσαρμόζονται στις ιδιαιτερότητες των κειμένων αυτών. Μία βασική προσέγγιση η οποία υποστηρίζει αποτελεσματικά την παρουσίαση των πηγών των ανθρωπιστικών επιστημών στο νέο ψηφιακό περιβάλλον, είναι η **κωδικοποίηση κειμένων** (text encoding, mark up text).

Η κωδικοποίηση χρησιμοποιείται σε διάφορα επιστημονικά πεδία. Σε κάθε περίπτωση δηλώνει μία διαφορετική διαδικασία η οποία διαμορφώνεται και προσαρμόζεται στο γνωστικό αντικείμενο και τις απαιτήσεις της κάθε επιστήμης. Στις ανθρωπιστικές επιστήμες η διαδικασία αυτή σχετίζεται με την κωδικοποίηση πηγών με μορφή κειμένου. Αναφέρεται σε μία διαδικασία που στηρίζεται σε προκαθορισμένα πρότυπα και αρχές. Υποδηλώνει τη δημιουργία, τη δομημένη εμφάνιση και την επεξεργασία του κειμένου στο ψηφιακό περιβάλλον, έτσι ώστε να είναι συμβατό με όλες τις υπάρχουσες, αλλά και τις μελλοντικές τεχνολογίες. Στόχος της είναι η δημιουργία ενός ηλεκτρονικού κειμένου σε μια μορφή η οποία επιτρέπει την ανάγνωσή του και ανταποκρίνεται στις πληροφοριακές ανάγκες του εκάστοτε χρήστη.

Η κωδικοποίηση δεν αποτελεί επίτευγμα και αποκλειστικότητα της πληροφορικής. Για παράδειγμα, τόσο τα σημεία στίξης όσο και τα τυπογραφικά στοιχεία δεν είναι τίποτε άλλο από μορφές κωδικοποίησης κειμένων. Τα σημεία στίξης τοποθετούνται στο εσωτερικό του κειμένου και διακρίνουν τα δομικά χαρακτηριστικά του π.χ. χωρισμός λέξεων και δημιουργία παραγράφων. Άλλη μορφή κωδικοποίησης είναι η ανακάλυψη της τυπογραφίας (σελιδοποίηση, στοιχειοθέτηση, μορφοποίηση κ.ά.), στο πλαίσιο της οποίας και πρωτοπαρουσιάστηκε ο όρος «κωδικοποίηση» ως δηλωτικός της διαδικασίας εισαγωγής τυπογραφικών στοιχείων. Τα στοιχεία αυτά διευκολύνουν τη μελέτη, την κατανόηση και τη μετάδοση των πληροφοριών που εμπεριέχονται στο γραπτό έργο.

Η διαδικασία κωδικοποίησης διακρίνεται σε δύο κατηγορίες:

- στην ειδική κωδικοποίηση (specific mark up), που ασχολείται με την παρουσίαση ενός τεκμηρίου στο χρήστη, και
- στη γενικευμένη κωδικοποίηση (generalized mark up), που η έμφαση δίνεται πρωτίστως στη δομή του κειμένου και δευτερευόντως στη μορφοποίηση και στον τρόπο εμφάνισης του.

Η **ψηφιοποίηση** και η **κωδικοποίηση** είναι δύο διαφορετικές έννοιες οι οποίες δεν πρέπει να συγχέονται, ούτε να ταυτίζονται. Η ψηφιοποίηση, αφορά μια διαδικασία μετατροπής ενός τεκμηρίου από αναλογική σε ψηφιακή μορφή, ενώ ως κωδικοποίηση κειμένου ορίζεται «η διαδικασία διάκρισης δομικών και θεματικών

(semantics) χαρακτηριστικών του κειμένου με βάση κάποιους κανόνες» (Γεργατσούλης, Παπαθεοδώρου, 2004). Στόχος της κωδικοποίησης είναι η δημιουργία ενός κειμένου, το οποίο θα είναι ευανάγνωστο και θα ανταποκρίνεται στις πληροφοριακές ανάγκες του εκάστοτε χρήστη.

Πρέπει να σημειωθεί πως η κωδικοποίηση που γίνεται σε ένα κείμενο με το πρότυπο TEI χαρακτηρίζεται ως γενικευμένη (generalized mark up), καθώς δίνει έμφαση στη δομή του τεκμηρίου. Η κωδικοποίηση δεν αφορά μόνο πηγές σε μορφή κείμενου, αλλά αφορά εξίσου την παραγωγή και τη διακίνηση υλικού μέσω του Διαδικτύου, συμβάλλοντας καθοριστικά στην ανάπτυξη της ηλεκτρονικής δημοσίευσης (electronic Publishing). Κατά τη διάρκεια της τελευταίας δεκαπενταετίας δημιουργήθηκε ένα εξειδικευμένο πρότυπο που προσπαθεί να καλύψει τις ανάγκες της κωδικοποίησης κειμένων. Το πρότυπο αυτό ονομάζεται **TEI** (**T**ext **E**ncoding **I**nitiative – Πρωτοβουλία Κωδικοποίησης Κειμένου).

Text Encoding Initiative

Το **TEI** αφορά μια διεθνή προσπάθεια, που ξεκίνησε το 1987 και ασχολείται με την ανάπτυξη οδηγιών για την προετοιμασία και την ανταλλαγή ηλεκτρονικών κειμένων στο πλαίσιο της επιστημονικής έρευνας. Το TEI έχει δημιουργήσει ένα σύνολο SGML DTDs για την κωδικοποίηση των κειμένων που αφορούν τις ανθρωπιστικές και τις κοινωνικές επιστήμες. Ενώ αρχικά δημιουργήθηκε για να ικανοποιεί τις ανάγκες ανταλλαγής δεδομένων, έχει εξελιχθεί πλέον σε ένα ισχυρό μέσο αναζήτησης, ευρετηρίασης και αποθήκευσης πληροφοριών, το οποίο έχει αξιοποιηθεί διεθνώς από βιβλιοθήκες, μουσεία και υπηρεσίες πληροφόρησης, με αποτέλεσμα να έχει αναχθεί σε διεθνώς διαδεδομένο πρότυπο κωδικοποίησης.

Το πρότυπο TEI είναι ουσιαστικά ένα λεξιλόγιο, αρχικά συμβατό με τη γλώσσα **SGML** (**S**tandard **G**eneralized **M**arkup **L**anguage) και έπειτα με την **XML** (**Ex**tensible **M**arkup **L**anguage), το οποίο εκφράζεται μέσα από ένα DTD (**D**ocument **T**ype **D**efinition - Δήλωση Τύπων Τεκμηρίου). Η SGML και η XML, αν και συνήθως χαρακτηρίζονται ως γλώσσες, στην πραγματικότητα είναι μεταγλώσσες (metalanguages), αφού μπορούν να χρησιμοποιηθούν για τον ορισμό νέων λεξιλογίων και γλωσσών σήμανσης. Χάρη στην ιδιότητά τους αυτή μπορούν να συμβάλλουν καθοριστικά στη δημιουργία προτύπων κωδικοποίησης δεδομένων. Το DTD, με την σειρά του, περιλαμβάνει ένα λεξιλόγιο, καθώς και τους κανόνες που διέπουν τα στοιχεία ενός XML τεκμηρίου, όπως τα στοιχεία, τη σειρά εμφάνισης των στοιχείων, το πλήθος εμφανίσεων τους κ.α. Σε αυτές τις ιδιότητες στηρίχθηκε και η ανάπτυξη του προτύπου TEI, το οποίο χρησιμοποιείται για την κωδικοποίηση λογοτεχνικών και γλωσσολογικών κειμένων, παράγοντας ηλεκτρονικά κείμενα, τα οποία είναι εκμεταλλεύσιμα και συμβατά με οποιοδήποτε υπολογιστικό σύστημα χωρίς την παραμικρή απώλεια πληροφορίας.[1]

Δομή τεκμηρίου TEI

Ένα κείμενο, το οποίο έχει κωδικοποιηθεί με το πρότυπο TEI αποτελείται από τα παρακάτω βασικά στοιχεία:[2]

---

[1] http://www.skaldic.arts.usyd.edu.au/docs/electro.guide/original.html (ημερομηνία πρόσβασης 25.7.2005)

[2] http://bistro.northwestern.edu/AnaServer?/tei+0+frame.anv/ (ημερομηνία πρόσβασης 25.7.2005)

- Το στοιχείο **teiHeader**, το οποίο δεν αποτελεί μέρος του περιεχομένου του κειμένου, αλλά επεξεργάζεται τις βιβλιογραφικές πληροφορίες του (τίτλος, συγγραφές, εκδότης, κτλ.), δηλαδή παράγει τα μεταδεδομένα (metadata).
- Το στοιχείο **text**, περιλαμβάνει το περιεχόμενο του κωδικοποιημένου κειμένου και αποτελείται από τα παρακάτω επιμέρους στοιχεία:
  1. Το στοιχείο **front** (προαιρετικό στοιχείο), το οποίο περιέχει το προκαταρκτικό περιεχόμενο που προηγείται του κυρίως κειμένου (επικεφαλίδες, σελίδες τίτλων, πρόλογοι, λίστες ρόλων, κτλ.)
  2. Το στοιχείο **body** (υποχρεωτικό στοιχείο), το οποίο περιέχει το σώμα ενός μοναδικού κειμένου. Δεν συμπεριλαμβάνονται τα στοιχεία front και back. Πρόκειται για ένα ιδιαίτερα πολύπλοκο στοιχείο που διαμορφώνεται ανάλογα με το είδος του κωδικοποιημένου κειμένου (π.χ. θεατρικό, πεζό κτλ).
  3. Το στοιχείο **back** (προαιρετικό στοιχείο), το οποίο περιλαμβάνει το περιεχόμενο που εμφανίζεται μετά το κυρίως κείμενο π.χ. παραρτήματα, γλωσσάρια, βιβλιογραφία κ.ά.

Μορφή κωδικοποιημένου κειμένου

```
<TEI.2>
<teiHeader><!--περιεχόμενο του στοιχείου--></teiHeader>
<text>
   <front>
      <!--εξώφυλλο (front matter)-->
   </front>
   <body>
  <!--κυρίως κείμενο-->
   </body>
   <back>
 <!--οπισθόφυλλο (back matter)-->
</back>
</text>
 </TEI.2>
```

Εκδόσεις του TEI

Μέχρι σήμερα έχουν κυκλοφορήσει πέντε εκδόσεις του προτύπου TEI. Οι εκδόσεις αυτές συνοδεύονται από έναν πολυσέλιδο οδηγό (TEI Guidelines), ο οποίος περιγράφει αναλυτικά την εκάστοτε έκδοση του προτύπου. Η πρώτη έκδοσή του κυκλοφόρησε τον Ιούνιο του 1990 και ήταν συμβατή με τη γλώσσα SGML. Η έκδοση αυτή είναι γνωστή με τα αρχικά **TEI P1** (το αρχικό **P** δηλώνει τη λέξη "**P**roposal"/πρόταση). Το 1992 εκδόθηκε η δεύτερη έκδοση του προτύπου, η **TEI P2** – σε γλώσσα SGML - η οποία περιλαμβάνει νέο υλικό καθώς και τις αλλαγές που έχουν γίνει σε σχέση με την έκδοση TEI P1. Το 1999 κυκλοφόρησε η τρίτη έκδοση, η **TEI P3**, η οποία εκφράζεται και πάλι σε γλώσσα SGML και είναι αρκετά βελτιωμένη συγκριτικά με τις προηγούμενες εκδόσεις. Το 2002 κυκλοφόρησε η τέταρτη έκδοση η **TEI P4**. Χαρακτηριστικό της έκδοση αυτής είναι η συμβατότητά της με την XML.

Τέλος, στις αρχές του 2005 κυκλοφόρησε και η τελευταία έκδοση, η **TEI P5**, η οποία επίσης είναι συμβατή με την XML και τις τεχνολογίες που υποστηρίζουν τη

γλώσσα αυτή. Η έκδοση αυτή πρόκειται να λύσει πρακτικά και θεωρητικά προβλήματα, καθώς το πρότυπο εκφράζεται μέσα από ένα Schema, το RelaxNG XML και όχι με ένα DTD. Η RELAX NG είναι μια απλή γλώσσα σχήματος για την XML, η οποία βασίζεται στην TRELAX και την TREX. Ένα σχήμα RELAX NG καθορίζει ένα πρότυπο σχέδιο (pattern) για τη δομή και το περιεχόμενο ενός τεκμηρίου XML, προσδιορίζοντας έτσι μια κατηγορία τεκμηρίων XML η οποία αποτελείται από όλα τα τεκμήρια τα οποία ταιριάζουν με το συγκεκριμένο πρότυπο σχέδιο. Το RELAX NG σχήμα, αυτό καθ' αυτό, είναι ένα τεκμήριο XML. Έτσι, είναι δυνατή η αξιοποίηση και η χρήση υποστηρικτικών τεχνολογιών της XML, παρέχοντας παράλληλα στο χρήστη περισσότερες δυνατότητες για την κωδικοποίηση ενός κειμένου.

Παρατηρούμε πως ενώ οι τρεις πρώτες εκδόσεις του TEI (P1, P2 και P3) εκφράζονται σε γλώσσα SGML, οι δύο τελευταίες (P4 και P5) είναι συμβατές με την XML. Ο πιο προφανής λόγος της χρήσης της XML αντί της SGML - μιας γλώσσας ιδιαίτερα διαδεδομένης με την οποία οι χρήστες παρουσιάζονται να έχουν ιδιαίτερη εξοικείωση - είναι οι δυνατότητες που προσφέρει η γλώσσα XML. Η XML είναι σχεδιασμένη με τέτοιο τρόπο, ώστε τα τμήματα (subtrees) ενός XML τεκμηρίου να μπορούν να αναλυθούν σωστά, ως μεμονωμένα και ανεξάρτητα τμήμα, ακόμα και αν έχουν αποσπαστεί από το κείμενο, από το οποίο προέρχονται. Συνεπώς δεν είναι, απαραίτητη η μεταφορά ολόκληρου του κειμένου στο Διαδίκτυο προκειμένου ο χρήστης να ασχοληθεί με συγκεκριμένα τμήματά του. Η ιδιότητα αυτή είναι ιδιαίτερα σημαντική, καθώς τα TEI τεκμήρια έχουν συνήθως μεγάλη έκταση. Έτσι, οι πληροφορίες ενός TEI τεκμηρίου μεταφέρονται απευθείας στον χρήστη με εύχρηστα και χαμηλού κόστους εργαλεία. Ένα ζήτημα που απασχόλησε το TEI Consortium, όταν κυκλοφόρησε η τέταρτη έκδοση του προτύπου, αφορούσε στη μεταφορά των τεκμηρίων TEI από την έκδοση SGML P3 στην XML P4. Λύση στο ζήτημα αυτό δόθηκε το 2002 με την ανάπτυξη της μεθόδου **μετανάστευσης** (migration).

Μετανάστευση

Αν και η μετανάστευση, εκτός των άλλων αποσκοπεί και στην ανανέωση των τεκμηρίων - όσον αφορά τη συντήρηση και την μελλοντική πρόσβασής τους - με τον έννοια που χρησιμοποιείται εδώ, αφορά μόνο τη μεταφορά (ή την μετατροπή) ψηφιακών αρχείων από μια προγενέστερη τεχνολογία ή ένα λογισμικό σε κάποιο νεώτερο.

Η μετανάστευση των τεκμηρίων TEI, από SGML σε XML, παρέχει πολλαπλά οφέλη. Με δεδομένη την ταχύτητα ανάπτυξης της τεχνολογίας και κατά συνέπεια των δυνατοτήτων που αυτή προσφέρει, οδηγούμαστε στην ανάγκη επανεξέτασης πολλών προγραμμάτων που λειτουργούν ήδη επί σειρά ετών με το ίδιο SGML DTD. Η μετανάστευση παρέχει μια μοναδική ευκαιρία να επανεξεταστούν οι πρακτικές των DTDs και της κωδικοποίησης που αναπτύχθηκαν σε ένα σύστημα βασισμένο στην SGML, τα οποία όμως δεν κρίνονται πλέον απαραίτητα στο περιβάλλον ενός συστήματος που υποστηρίζει την XML (Ruotolo, 2004).

Αφ' ενός, λόγω της χρήσης της XML ως του κυρίαρχου προτύπου, η μελλοντική ύπαρξη και υποστήριξη λογισμικού που βασίζεται στην SGML καθίσταται αμφισβητήσιμη και αφ' ετέρου, λόγω του γεγονότος ότι η XML συνοδεύεται από διάφορα σχετικά πρότυπα και προδιαγραφές - όπως τα σχήματα XPath, XSLT, XML, XPointer, XLink, και XQuery (Ruotolo, 2004), γίνεται αναγκαία η μελέτη σχεδίων και η λήψη αποφάσεων σχετικά με τη μετανάστευση των κωδικοποιημένων τεκμηρίων. Η μετανάστευση μπορεί να οδηγήσει σε μείωση του κόστους, καθώς κυκλοφορούν δωρεάν υψηλής ποιότητας XML εργαλεία. Παράλληλα, οι χρήστες

εξοικειώνονται όλοι και περισσότερο με την XML και έτσι περιορίζεται το κόστος σε εκπαίδευση και χρόνο. Τέλος, δίνεται και μια ευκαιρία να γίνουν έλεγχοι και να διορθωθούν λάθη και παραλήψεις στα ήδη κωδικοποιημένα τεκμήρια.

Στο πλαίσιο της ανάπτυξης του TEI, δημιουργήθηκε μια ομάδα εργασίας η οποία στελεχώθηκε από αντιπροσώπους ιδρυμάτων και οργανισμών που κατέχουν μεγάλα αποθετήρια τεκμηρίων σε SGML TEI, καθώς και από τεχνικούς εμπειρογνώμονες και συντάκτες τεκμηρίων TEI. Η ομάδα αυτή προσπαθεί να τεκμηριώσει τις μεθόδους και τα εργαλεία που είναι απαραίτητα για τη μετανάστευση των υπαρχόντων δεδομένων σε TEI XML.

Από την P3 στην P4

Η μετανάστευση δεν είναι μια τετριμμένη διαδικασία, ειδικά στις περιπτώσεις μεγάλων ποσοτήτων δεδομένων. Αποτελεί μία διαδικασία που ολοκληρώνεται σε τέσσερα βήματα (Strategic considerations…, 2004):

1. Μετατροπή των TEI P3-SGML τεκμηρίων σε TEI P4-XML. Η μετατροπή γίνεται με τη χρήση διαφόρων εργαλείων, όπως osx[3], n2x[4], Arbortext Epic[5], XMetaL[6],κ.ά.
2. Έλεγχος στις ετικέτες που χρησιμοποιούνται για την κωδικοποίηση του κειμένου. Η ανάγκη αυτή προκύπτει από το γεγονός ότι στην XML είναι σημαντική η διάκριση κεφαλαίων-πεζών, ενώ στην SGML αυτό δεν είναι απαραίτητο.
3. Μορφοποίηση των κωδικοποιημένων κειμένων ώστε να είναι ευανάγνωστα.
4. Ανάπτυξη στρατηγικών ελέγχου και επίλυσης προβλημάτων που ίσως παρουσιαστούν κατά τη διαδικασία μετανάστευσης.

Ο βαθμός δυσκολίας της μετατροπής των τεκμηρίων από την έκδοση P3 (SGML) στην P4 (XML), εξαρτάται από το επίπεδο κωδικοποίησης, την ομοιογένεια των συλλογών και την χρήση κοινού ή διαφορετικού DTD, καθώς και την πιθανή ανάγκη δημιουργίας ενός νέου XML DTD. Κατά τη διαδικασία της μετανάστευσης θα πρέπει να λαμβάνονται υπ' όψιν και άλλοι παράγοντες όπως, το μέγεθος της συλλογής, ο χρόνος και το προσωπικό που θα ασχοληθεί με αυτή την εργασία. Αυτονόητο είναι ότι μεγάλες και πιο πολύπλοκες συλλογές απαιτούν περισσότερο χρόνο καθώς και ανθρώπινο δυναμικό. Επίσης ο φορέας υλοποίησης θα πρέπει να αποφασίσει αν παράλληλα με τη διαδικασία μετανάστευσης θα παράγει νέα κείμενα σε XML ή θα σταματήσει την παραγωγή μέχρι να ολοκληρωθεί η διαδικασία αυτή (Strategic considerations…, 2004)

Συμπεράσματα

Η μετανάστευση των τεκμηρίων TEI από την έκδοση P3 (SGML) στην P4 (XML) είναι σχετικά απλή και έχουν ήδη αναπτυχθεί ειδικά scripts (σειρές εντολών) για την υποστήριξή της. Η μετανάστευση από την P3 στην P4 κρίνεται απαραίτητη, καθώς η άμεση μετάβαση από την P3 στην P5 θα είναι εξαιρετικά δύσκολη. Η τελευταία έκδοση, TEI P5, υποστηρίζει την XML χωρίς να εξασφαλίζεται παράλληλα η συμβατότητά της με την P4. Ταυτόχρονα, το TEI Consortium, δεν σκοπεύει να

---

[3]osx: πρόκειται για το πιο δημοφιλές εργαλείο που χρησιμεύει για τη μετάβαση από τη γλώσσα SGML στην XML. Προέρχεται από το εργαλείο sx.

[4]n2x: είναι ένα open source εργαλείο που χρησιμοποιείται για τη μετάβαση από τη γλώσσα SGML στην XML.

[5] Arbortext Epic: είναι ένας SGML editor που χρησιμοποιείται για την μετατροπή σε XML

[6] XMetaL: είναι ένας XML editor

προσφέρει μια τέτοια δυνατότητα, διότι η μετάβαση από την έκδοση P4 στην P5 θα είναι πιο εύκολη, αφού θα γίνεται μεταφορά από ένα XML DTD σε RelaxNG. Επομένως, η διαδικασία μετανάστευσης από την έκδοση P3 σε P4, θα γίνεται για πολλά χρόνια και κατά συνέπεια θα εξακολουθεί να χρησιμοποιείται παράλληλα και η τέταρτη έκδοση του TEI. Έτσι η μετανάστευση από την P3 στην P5, καθίσταται ουσιαστικά δυσκολότερη από ότι η μετανάστευση από την P3 στην P4. Επομένως η μετατροπή όλων των κειμένων από προηγούμενες εκδόσεις στην έκδοση TEI P4, είναι ο μόνος τρόπος με τον οποίο μπορεί να καταστεί δυνατή τη μελλοντική χρήση των κωδικοποιημένων κειμένων, όταν γίνει απαραίτητη η μετανάστευσή τους στην έκδοση P5. Αν και πρόκειται για ένα ζήτημα που έχει αρχίσει να μελετάται από το 2002, ουσιαστικά οι πρώτες δημόσιες ανακοινώσεις ξεκίνησαν την εμφάνισή τους μόλις κατά τη διάρκεια του 2004, γεγονός που συνεπάγεται πως υπάρχουν ακόμη ανοικτά ζητήματα που χρήζουν περαιτέρω ελέγχου και έρευνας.

Βιβλιογραφία

1. Christine Ruotolo, *Recommendations of the TEI Task Force on SGML to XML migration*, 2004 [**τεκμήριο** www, URL: http://www.hum.gu.se/allcach2004/AP/html/prop146.html, **ημερομηνία πρόσβασης**: 29.10.2005].
2. John Unsworth, *What is humanities computing and what is not?*, 2002 [**τεκμήριο** www, URL: http://computerphilologie.uni-muenchen.de/jg02/unsworth.html, **ημερομηνία πρόσβασης**: 6.6.2005].
3. Maciej Ogrodniczuk, From SGML to XML with TEI: automated conversion of a corpus of polish from P3 to P4 format, *Investigationes Linguisticae*, Dec. 2004 [**τεκμήριο** www, URL: http://www.staff.amu.edu.pl/~inveling/maciej_ogrodniczuk_inve11.pdf, **ημερομηνία πρόσβασης**: 29.10.2005].
4. Practical Guide to Migration of TEI Documents from SGML to XML, *TEI Consortium* [**τεκμήριο** www, URL: http://www.tei-c.org/Activities/MI/miw03.html **ημερομηνία πρόσβασης**: 29.10.2005].
5. Strategic considerations in migration of TEI documents from SGML to XM, *TEI Consortium*, 2004 [**τεκμήριο** www, URL: http://www.tei-c.org/Activities/MI/miw02.html, **ημερομηνία πρόσβασης**: 9.8.2005].
6. Technical Checklist for TEI/SGML documents, *TEI Consortium*, 2002 [**τεκμήριο** www, URL: http://www.tei-c.org/Activities/MI/miw04.html, **ημερομηνία πρόσβασης**: 28.10.2005].
7. TEI SGML to XML migration. Introduction and workflow recommendations. Second draft, 2003, *TEI Consortium* [**τεκμήριο** www, URL: http://www.tei-c.org/Activities/MI/miw03d.html, **ημερομηνία πρόσβασης**: 2.7.2005].
8. TEI Task Force on SGML to XML Migration, *TEI Consortium*, 2004 [**τεκμήριο** www, URL: http://www.tei-c.org/Activities/MI/, **ημερομηνία πρόσβασης**: 28.10.2005].
9. **Μανόλης Γεργατσούλης, Χρήστος Παπαθεοδώρου**, *Πρότυπα κωδικοποίησης ΙΙ: σημειώσεις για το μάθημα Πρότυπα κωδικοποίησης* **ΙΙ, Ιόνιο Πανεπιστήμιο, ΤΑΒ**, 2004.