Συνιστώμενα μορφότυπα αρχείων CLARIN:EL

Συστάσεις για την επιλογή μορφότυπων αρχείων που διασφαλίζουν μακροπρόθεσμα την προσβασιμότητα και διαλειτουργικότητα των δεδομένων

Στο παρόν έγγραφο παρουσιάζονται τα μορφότυπα αρχείων που συνιστώνται για την μεταφόρτωση δεδομένων στο CLARIN:EL. Οι οδηγίες έχουν συνταχθεί από την τεχνική ομάδα της Υποδομής CLARIN:EL (technical-helpdesk@clarin.gr), στην οποία μπορείτε να απευθυνθείτε για οποιαδήποτε ερώτηση ή παρατήρηση.

Μορφότυπα αρχείων που υπηρετούν την Πολιτική Ψηφιακής Διατήρησης

Κατά τον καθορισμό των συνιστώμενων μορφότυπων αρχείων, η ομάδα του CLARIN:EL συνυπολόγισε τους ακόλουθους παράγοντες που επιτρέπουν στα δεδομένα να παραμένουν προσβάσιμα και να μπορούν να επαναχρησιμοποιηθούν σε βάθος χρόνου:

- Για την **επεξεργασιμότητα** (processability) λήφθηκαν υπόψιν:
    - η καταλληλότητα του μορφότυπου για τον τύπο του πόρου ή/και τον τύπο της επεξεργασίας, και
    - η συμβατότητά του με τις ροές εργασίας CLARIN:EL (τα κειμενικά δεδομένα, προκειμένου να είναι επεξεργάσιμα, πρέπει να είναι σε ένα από τα μορφότυπα που επεξεργάζονται οι ροές εργασίας).
- Για την **διατήρηση** (preservation) λήφθηκαν υπόψιν:
    - η καταλληλότητα του μορφότυπου για έρευνα, όπως έχει οριστεί από τις σχετικές κοινότητες,
    - ο βαθμός διάδοσής του (προτιμήθηκαν ευρέως χρησιμοποιούμενα μορφότυπα, που δεν έχουν καταργηθεί, γνωστά στις καθορισμένες κοινότητες),

- ο τρόπος διάθεσης (επιλέχθηκαν μορφότυπα ανοικτού έναντι κλειστού, ιδιόκτητου, κώδικα), και
- το εάν το μορφότυπο εφαρμόζει συμπίεση με ή χωρίς απώλειες.

Η πολιτική, που βασίζεται στα όσα προαναφέρθηκαν, ανταποκρίνεται στην αποστολή του CLARIN:EL να συλλέγει, να διατηρεί και να διανέμει ψηφιακούς γλωσσικούς πόρους και υπηρεσίες γλωσσικής επεξεργασίας για την υποστήριξη ερευνητών, ακαδημαϊκών, φοιτητών, επαγγελματιών του γλωσσικού τομέα, επιστημόνων καθώς και του ευρύτερου κοινού. Προκειμένου να δημιουργηθούν οι κατάλληλες συστάσεις για τα επιμέρους μορφότυπα αρχείων ή να αποφασιστεί η καταλληλότητά τους για συγκεκριμένα είδη ερευνητικών δραστηριοτήτων/τύπων, πρέπει να εξεταστεί ο σκοπός για τον οποίο προορίζονται. Για παράδειγμα, ενώ το μορφότυπο PDF/A έχει αναπτυχθεί για την χωρίς προβλήματα μακροπρόθεσμη αρχειοθέτηση και αποτελεί εξαιρετική επιλογή μορφότυπου για τεκμηρίωση, είναι αναμφίβολα *ακατάλληλο* για κειμενικά δεδομένα που προορίζονται για γλωσσική επεξεργασία. Επομένως, με βάση τους τύπους πόρων που εμπίπτουν στο πεδίο εφαρμογής των χρηστών του CLARIN:EL και των υπηρεσιών που προσφέρονται/υποστηρίζονται, η ομάδα του CLARIN:EL διακρίνει τις ακόλουθες κατηγορίες, ως προς τους ψηφιακούς γλωσσικούς πόρους, για τις οποίες παρέχονται συγκεκριμένες συστάσεις:

- **CLARIN:EL processable data** (επεξεργάσιμα δεδομένα στην υποδομή CLARIN:EL): πρόκειται για κειμενικά δεδομένα που μπορούν να τροφοδοτήσουν τις ροές εργασίας του CLARIN:EL,
- **Textual Data** (κειμενικά δεδομένα): γραπτό μη δομημένο/απλό κείμενο (plain text) ή δομημένο (π.χ. HTML) χωρίς γλωσσική ή άλλη σήμανση που προστίθεται για ερευνητικούς σκοπούς (*μη επεξεργάσιμα* από τις ροές εργασίας CLARIN:EL),
- **Text Annotation** (κειμενική επισημείωση): επισημειώσεις κειμενικών δεδομένων της γλώσσας προέλευσης, με το αρχικό κείμενο να συμπεριλαμβάνεται ή να υφίσταται ως ξεχωριστό έγγραφο (stand-off document),

- **Language Description** (γλωσσική περιγραφή): δεδομένα που περιγράφουν μια γλώσσα ή κάποια πτυχή/πτυχές μιας γλώσσας μέσω συστηματικής τεκμηρίωσης γλωσσικών δομών (γραμματικές, μοντέλα μηχανικής μάθησης (ML), μοντέλα N-γραμμάτων),
- **Lexical/Conceptual Resource** (λεξιλογικός/εννοιολογικός πόρος): πόρος οργανωμένος με βάση λεξιλογικές ή εννοιολογικές καταχωρίσεις (λεξιλογικά στοιχεία, όρους, έννοιες κ.λπ.) με συμπληρωματικές πληροφορίες (π.χ. μορφολογικές, σημασιολογικές, στατιστικές πληροφορίες κλπ.),
- **Image data** (δεδομένα εικόνας): ψηφιοποιημένες εικόνες αναλογικών πηγών δεδομένων γραπτού λόγου για ερευνητικούς σκοπούς (π.χ. σαρώσεις χειρογράφων, φωτογραφίες επιγραφών) ή δισδιάστατες εικόνες ή σχήματα που διανέμονται με τα σχετικά κειμενικά δεδομένα για ανάλυση επεξεργασίας φυσικής γλώσσας, NLP, (π.χ. ιατρικές εικόνες, *δεδομένα εικόνας*, συνοδευόμενες από ακτινολογικές γνωματεύσεις, *κειμενικά δεδομένα*),
- **Audio data** (ηχητικά δεδομένα): ηχογραφήσεις που προσφέρουν δεδομένα προφορικού λόγου για ερευνητικούς σκοπούς (π.χ. αρχεία ήχου με την προφορά λέξεων για ένα λεξικό, ηχογραφημένες συνεντεύξεις, ραδιοφωνικές εκπομπές κλπ.),
- **Video data** (δεδομένα βίντεο): βιντεοσκοπήσεις που προσφέρουν πολυτροπικά ή δεδομένα νοηματικής γλώσσας για ερευνητικούς σκοπούς.

Συστάσεις μορφότυπων

Προτιμώνται τα μορφότυπα που πληρούν τα κριτήρια της Πολιτικής Ψηφιακής Διατήρησης που προαναφέρθηκαν. Ωστόσο, γίνονται δεκτά και άλλα μορφότυπα, ως *επίπεδο πρώτης εισόδου*, που προτείνεται να μετατραπούν στα συνιστώμενα.

Ως εκ τούτου, τα μορφότυπα αρχείων κατηγοριοποιούνται σε δύο επίπεδα διατήρησης (**recommended**/συνιστώμενα, **acceptable**/αποδεκτά) πάντα στο πλαίσιο της κάθε περίπτωσης. Ο κατάλογος των αποδεκτών μορφότυπων δεν είναι εξαντλητικός, ιδίως

στην περίπτωση της κειμενικής επισημείωσης (text annotation), αλλά μάλλον ενδεικτικός, ενώ προτείνεται και η μετατροπή των αποδεκτών μορφότυπων σε συνιστώμενα.

	Συνιστώμενα	**Αποδεκτά**
CLARIN:EL processable data	**Monolingual textual data**: plain text **Monolingual encoded data**: XCES-ILSP variant (XML based format compliant with the XCES model for corpora) **Bi-/Multilingual encoded data**: TMX (XML based format for aligned data), MOSES (text-based format for parallel data)	
Textual Data	**File Formats**: plain text **Formatted/Encoded**: ODT, DOCX, PDF/A, HTML, Latex, TeX, MOSES	PDF, SGML, Rich Text Format (.rtf), Microsoft Word (.doc, .docx), PostScript

Text Annotation	**File Formats**: XML, XMI, CSV, TSV, RDF (all serialisation formats RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), JSON **Models**: XCES for corpora and structural annotation, TEI for structural and linguistic annotation, GrAF linguistic annotation, TMX for aligned, GATE linguistic annotation, CoNLL family (CoNLL-U, CoNLL-2000, CoNLL-2002, CoNLL-2003, CoNLL-2006, CoNLL-2008, CoNLL-2009, CoNLL-2012) for linguistic annotation, NIF linguistic annotation for RDF data, WARC for web crawled data	**File Formats**: SGML, Plain Text, Microsoft Excel (.xlsx, .xls), ELLOGON
Language Description	**ML Model**: H5, ProtoBuf, ONNX, PMML, Pickle, MLeap, YAML, JSON **N-gram model**: ARPA	
Lexical / Conceptual Resource	**File Formats**: XML, CSV, TSV, RDF (RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), OWL **Models**: LMF for lexica, OWL for ontologies, SKOS for thesauri, OntoLex-Lemon for lexica, TBX for terminological data	Microsoft Excel (.xlsx, .xls), Plain Text, SQL

Image data	**All images**: TIFF, SVG, JPEG 2000, PNG, GIF **Scanned images**: PDF/A	JPEG, BMP, Photoshop, NifTi, FlashPix, PDF
Audio data	WAV, AIFF, FLAC	MP3, MPEG, Windows Media Audio
Video data	AVI	MPEG-4, RealNetworks 'Real Video', Windows Media Video, Flash Video, QuickTime Video