

Εισήγηση προς το Τομεακό Επιστημονικό Συμβούλιο (ΤΕΣ) για τις Κοινωνικές και στις Ανθρωπιστικές Επιστήμες (ΚΑΕ)

(draft, 27-4-2021)

Μεγαδεδομένα (Big Data)



Υπο-ομάδα ΤΕΣ, Χρήστος Καμπόλης - Δημήτρης Μπάλλας - Μίλτος Πechλιβάνος

Έργο της 2ης Υπο-ομάδας: Προτάσεις σχετικά με τη δημιουργία σε εθνικό επίπεδο δομών που θα συλλέξουν ή/και θα αξιοποιήσουν υπάρχοντα μεγαδεδομένα (Big Data) στις κοινωνικές (οικονομικά, ψυχολογία, κοινωνιολογία, κλπ.) και στις ανθρωπιστικές επιστήμες. Θέματα προς συζήτηση, μεταξύ άλλων, είναι η πρόσβαση και η σύνδεση σε βάσεις δεδομένων, προσωπικά δεδομένα, κλπ.

1. Τι είναι τα μεγαδεδομένα;

‘Τα μαζικά δεδομένα είναι σύνολα δεδομένων που έχουν συλλεχθεί και είναι τόσο μεγάλα σε όγκο και περίπλοκα που απαιτούν νέες τεχνολογίες, όπως η τεχνητή νοημοσύνη, για την επεξεργασία τους. Τα δεδομένα προέρχονται από πολλές διαφορετικές πηγές. Συχνά είναι δεδομένα του ίδιου τύπου, για παράδειγμα, δεδομένα GPS που προέρχονται από εκατομμύρια κινητά τηλέφωνα και χρησιμοποιούνται για τον περιορισμό της κυκλοφοριακής συμφόρησης. Ωστόσο μπορούμε να έχουμε και συνδυασμούς τύπων, όπως στην περίπτωση συλλογής δεδομένων από μητρώα υγείας και εφαρμογές εξυπηρέτησης ασθενών. Η τεχνολογία επιτρέπει τη συλλογή αυτών των δεδομένων με υψηλές ταχύτητες, σε σχεδόν πραγματικό χρόνο, και την ανάλυση τους για τη απόκτηση νέων πληροφοριών.’

Ευρωπαϊκό Κοινοβούλιο, 2021

(<https://www.europarl.europa.eu/news/el/headlines/society/20210211STO97614/mazika-dedomena-orismos-ofeli-kai-prokliseis-grafimata>)

Ειδικότερα ως προς τη συλλογή μεγαδεδομένων στο πλαίσιο των ερευνητικών ζητήσεων των ΚΑΕ θα πρέπει να υπάρξει μέριμνα για την υιοθέτηση κριτηρίων και διαδικασιών ποιοτικού ελέγχου ως προς τα 4 + 1 Vs, βάσει των οποίων ορίζεται και τυποποιείται ένα σύνολο δεδομένων ως Big Data: *Volume, Veracity, Velocity, Variety + Value*.

***Volume:** The main characteristic that makes data “big” is the sheer volume. It makes no sense to focus on minimum storage units because the total amount of information is growing exponentially every year. In 2010, Thomson Reuters estimated in its annual report that it believed the world was “awash with over 800 exabytes of data and growing.”*

***Variety:** Variety is one the most interesting developments in technology as more and more information is digitized. Traditional data types (structured data) include things on a bank statement like date, amount, and time. These are things that fit neatly in a relational database.*

Structured data is augmented by unstructured data, which is where things like Twitter feeds, audio files, MRI images, web pages, web logs are put — anything that can be captured and stored but doesn’t have a meta model (a set of rules to frame a concept or idea — it defines a class of information and how to express it) that neatly defines it.

Unstructured data is a fundamental concept in big data. The best way to understand unstructured data is by comparing it to structured data. Think of structured data as data that is well defined in a set of rules. For example, money will always be numbers and have at least two decimal points; names are expressed as text; and dates follow a specific pattern.

With unstructured data, on the other hand, there are no rules. A picture, a voice recording, a tweet — they all can be different but express ideas and thoughts based on human understanding. One of the goals of big data is to use technology to take this unstructured data and make sense of it.

***Veracity:** Veracity refers to the trustworthiness of the data. Can the manager rely on the fact that the data is representative? Every good manager knows that there are inherent discrepancies in all the data collected.*

***Velocity:** Velocity is the frequency of incoming data that needs to be processed. Think about how many SMS messages, Facebook status updates, or credit card swipes are being sent on a particular telecom carrier every minute of every day, and you’ll have a good appreciation of velocity. A streaming application like Amazon Web Services Kinesis is an example of an application that handles the velocity of data.*

***Value:** It may seem painfully obvious to some, but a real objective is critical to this mashup of the four V’s. Will the insights you gather from analysis create a new product line, a cross-sell opportunity, or a cost-cutting measure? Or will your data analysis lead to the discovery of a critical causal effect that results in a cure to a disease? ’*

Jason Williamson, *Getting a Big Data Job For Dummies*, 2014

(Για το πώς εξειδικεύονται αυτοί οι γενικού τύπου ορισμοί στην έρευνα στις ΚΑΕ, βλ. για παράδειγμα: Jochen Tiepmar, *Big Data and Digital Humanities*, στο <https://publikationen.bibliothek.kit.edu/1000088937>)

2. Υπάρχουσες δομές μεγαδεδομένων για τις Κοινωνικές και Ανθρωπιστικές Επιστήμες

- Ελληνική Στατιστική Αρχή (<https://www.statistics.gr/>)
- Open access data διαθέσιμα από διεθνείς οργανισμούς, κρατικές και ιδιωτικές δομές στο εξωτερικό (π.χ. google trends, twitter, facebook, <https://www.data.gov/> κ.ά.)
- Μεγαδεδομένα που δημιουργούνται κυρίως από καθημερινές δραστηριότητες πολιτών, επιχειρήσεων και διαφόρων οργανισμών - ενδεικτικά παραδείγματα:
 - ο Ηλεκτρονικές συναλλαγές με πιστωτικές/χρεωστικές κάρτες (με πιθανές ενσωματωμένες πληροφορίες όπως ηλικία, φύλο, διεύθυνση χρηστών και καταστημάτων) - π.χ. supermarket data (including loyalty card related data)
 - ο Κενές θέσεις εργασίας (job vacancies) αγγελίες/διαφημίσεις /websites εταιριών κατά οικονομική δραστηριότητα (Πιερράκου και Μπισιώτη, 2016 - ΕΛΣΤΑΤ) - ειδικότητα, δεξιότητες εργαζομένων κ.ά.
(https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPB_Deliverable_B2_Methodological_framework_V1_2020_03_18.pdf)
 - ο Εμπορικές συναλλαγές (με πιθανά λεπτομερείς πληροφορίες σε σχέση με τον τύπο των συναλλαγών κατά οικονομική δραστηριότητα, με χωρικές πληροφορίες)
 - ο Δεδομένα σε σχέση με δίκτυα συγκοινωνιών (π.χ. δημόσιες συγκοινωνίες, τηλεματική, δρόμοι)
 - ο Κατανάλωση ενέργειας, βάση καταναλωτών ΔΕΗ, ενεργειακά πιστοποιητικά (που θα μπορούσαν να διατεθούν σε γεωγραφική κλίμακα οικοδομικού τετραγώνου) – πιθανές εφαρμογές στις Κοινωνικές Επιστήμες σε σχέση με ανάλυση ενεργειακής φτώχειας, κλιματική αλλαγή, οικολογικό αποτύπωμα
 - ο Φορολογικές Δηλώσεις – πιθανές εφαρμογές στις Κοινωνικές Επιστήμες σε σχέση με ανισότητες (και πιθανή σύνδεση με ενεργειακή φτώχεια, ανισότητες σε σχέση με education attainment)
 - ο Βάση δεδομένων με χρονοσειρά από βαθμολογικές επιδόσεις από σχολεία και αποτελέσματα πανελληνίων (με στοιχεία διεύθυνσης κατοικίας)
 - ο Υγειονομικά δεδομένα (νοσοκομεία, ιδιωτικοί φορείς παροχής υπηρεσιών υγείας)
- Εθνική υποδομή υποστήριξης ψηφιακών πόρων ΑΠΟΛΛΩΝΙΣ (<https://apollonis-infrastructure.gr>)

Με την εθνική υποδομή ΑΠΟΛΛΩΝΙΣ, που εντάσσεται στη Δράση «Ενίσχυση των Υποδομών Έρευνας και Καινοτομίας» και χρηματοδοτείται από το Επιχειρησιακό Πρόγραμμα «Ανταγωνιστικότητα, Επιχειρηματικότητα και Καινοτομία» στο πλαίσιο του ΕΣΠΑ 2014-2020, υποστηρίζονται και προωθούνται οι ψηφιακές ανθρωπιστικές επιστήμες και τέχνες, και η γλωσσική τεχνολογία και καινοτομία στην Ελλάδα. Η ΑΠΟΛΛΩΝΙΣ δημιουργήθηκε από τη σύμπραξη του Εθνικού Δικτύου Γλωσσικής Τεχνολογίας [clarin:el](#) και του Εθνικού Δικτύου Ψηφιακών Υποδομών για τις Ανθρωπιστικές Επιστήμες [DARIAH-GR/ΔΥΑΣ](#), τα οποία είναι συνιστώσες των αντίστοιχων ευρωπαϊκών υποδομών [CLARIN](#) και [DARIAH](#).

- Μητρώα Πόρων Ανθρωπιστικών Επιστημών στην Ελλάδα (<http://registries.dyas-net.gr/gr>)
Μέσω του εργαλείου ο χρήστης ερευνητής ή μελετητής των ανθρωπιστικών επιστημών και των τεχνών μπορεί να αναζητήσει πληροφορίες τόσο για φορείς (οργανισμούς ή ιδιώτες) της Ελλάδας, όσο και για τις συλλογές, φυσικές και ψηφιακές, που κατέχουν ή διαχειρίζονται.
- Ηλεκτρονικά αποθετήρια ψηφιακών πόρων - ενδεικτικά και μόνο παραδείγματα:
 - Ψηφιακή Βιβλιοθήκη του Εθνικού Κέντρου Τεκμηρίωσης (<http://www.ekt.gr/library>)
 - Ψηφιακές συλλογές του Εθνικού Ιδρύματος Ερευνών (<http://pandektis.ekt.gr/pandektis/>)
 - Ψηφιακές συλλογές Πανεπιστημιακών (ή άλλων) βιβλιοθηκών, λ.χ. η ιδιαιτέρως σημαντική για τις νεοελληνικές σπουδές του Πανεπιστημίου Κρήτης (<https://anemi.lib.uoc.gr/>)
 - Ψηφιοποιήσεις λογοτεχνικών και ιστορικών περιοδικών (βλ. το ευρετήριο στην Εθνική Βιβλιοθήκη: <https://www.nlg.gr/static-page/eyretirio-psifiopoiimenon-periodikon/>)
 - Συλλογές της Πύλης του Κέντρου Ελληνικής Γλώσσας (<https://www.greek-language.gr/greekLang/index.html>)

Πρβ. για το status quo των Digital Humanities στην Ελλάδα την παλαιότερη σύνοψη της Ελένης Γαρδίκια-Κατσιαδάκη:
(<https://www.hsozkult.de/searching/id/diskussionen-2401?title=the-status-quo-of-digital-humanities-in-greece&q=Katsiadakis&fq=&sort=newestPublished&total=5&recno=1&subType=fdd&ebate&title=the-status-quo-of-digital-humanities-in-greece&q=Katsiadakis&fq=&sort=newestPublished&total=5&recno=1&subType=fdd&ebate>)

3. Προτάσεις

- Η Υπο-ομάδα κρίνει για την υλοποίηση του έργου της ως σημαντικό τον συντονισμό και τις πιθανές συνεργασίες του ΤΕΣ ΚΑΕ με το ΤΕΣ Πληροφορικής αλλά και με το οριζόντιο ΤΕΣ Πολιτικής δεδομένων και Τεχνητής Νοημοσύνης.
- Η Υπο-ομάδα κρίνει ως αναγκαία για τη χαρτογράφηση του ιδιαιτέρως ποικιλόμορφου τοπίου των Ψηφιακών ΚΑΕ και τη διαμόρφωση προτάσεων συντονισμένης πολιτικής από το ΕΣΕΤΕΚ τη διοργάνωση με ευθύνη του ΤΕΣ μιας διημερίδας εργασίας εντός του χειμερινού εξαμήνου με την πρόσκληση φορέων που δραστηριοποιούνται στο πρόβλημα Big Data και Κοινωνικές/ανθρωπιστικές επιστήμες (ΕΛΣΤΑΤ, ΑΠΟΛΛΩΝΙΣ κ.ά.) και φορέων του δημόσιου και ιδιωτικού τομέα (π.χ. ΣΕΒ, ΟΤΕ, ΔΕΗ). Προτείνουμε ως βάση συζήτησης αυτών των συναντήσεων ερωτήματα όπως: *Ποιά είναι η μεθοδολογία της ΕΛΣΤΑΤ για την συλλογή μεγαδεδομένων; Υπάρχει συνεργασία με φορείς του ιδιωτικού και δημόσιου τομέα; Υπάρχει προγραμματισμός για τις κοινωνικές επιστήμες για*

πρωτοβουλίες αντίστοιχες της ΑΠΟΛΛΩΝΙΣ; Ποια (πρέπει να) είναι τα επόμενα βήματα για τη βελτιστοποίηση των ψηφιακών εργαλείων έρευνας και καινοτομίας στις ανθρωπιστικές επιστήμες; Ποιά (πρέπει να) είναι η στρατηγική για την τεχνογνωσία και υποδομές ανάλυσης και δυνατότητες συνεργασίας ιδιωτικού/δημόσιου τομέα (π.χ. τεχνητή νοημοσύνη , cloud computing)

- Η Υπο-ομάδα κρίνει ως αναγκαία τη συνεργασία δημόσιου και ιδιωτικού τομέα για τη χρηματοδότηση προγράμματος υποτροφιών νέων (μετα)διδακτόρων σε σχέση με μεγαδεδομένα/data science στα πρότυπα αναλογων προγραμμάτων όπως του Βρετανικού Economic and Social Research Council (<https://esrc.ukri.org/collaboration/postgraduate-collaboration/>)