

Καθαρισμός δεδομένων

Μαρία Ηλβανίδου

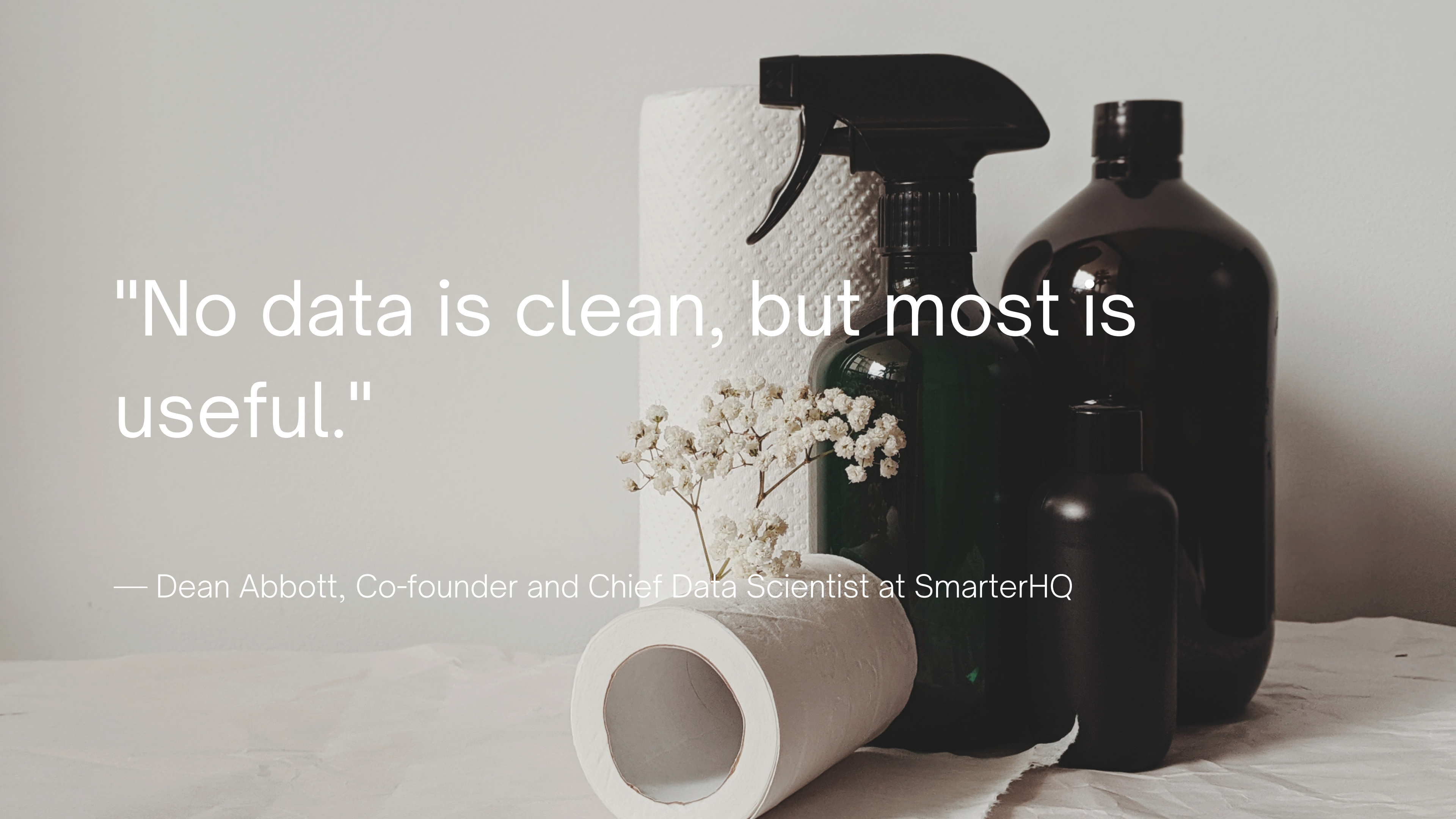
Επιστημονική Συνεργάτιδα | ΕΚ Αθηνά και ΟΠΑ

απολ^λωνίς

ΨΗΦΙΑΚΟ ΤΟΠΙΟ

ΑΘΗΝΑ Έρευνα & Καινοτομία
Τεχνολογίες Πληροφορίας

ΕΛΙΔΕΚ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας



"No data is clean, but most is useful."

— Dean Abbott, Co-founder and Chief Data Scientist at SmarterHQ

Για ποια δεδομένα μιλάμε

Όλα τα δεδομένα πιθανότατα χρειάζονται καθάρισμα πριν μπορέσουμε να τα αξιοποιήσουμε, προκειμένου να βελτιώσουμε την ποιότητά τους, ειδικά εάν προέρχονται από περισσότερες από μία πηγές.

Εδώ θα αναφερθούμε σε κειμενικά δεδομένα των Ανθρωπιστικών Επιστημών και τις ιδιαιτερότητές τους.

Τι είδους προβλήματα

Διαφορετική δομή

Πολλαπλές τιμές εντός του ίδιου πεδίου, τιμές που αναπτύσσονται σε περισσότερα πεδία και χρειάζεται να συμπτυχθούν σε ένα, κ.λπ.

Ελλείψεις στην τυποποίηση

Διαφορετική ορολογία, διαφορετική μορφή (π.χ. στην έκφραση της ημ/νίας), κ.λπ.

Σφάλματα

Ορθογραφικά, τυπογραφικά, πραγματολογικά, κ.λπ.

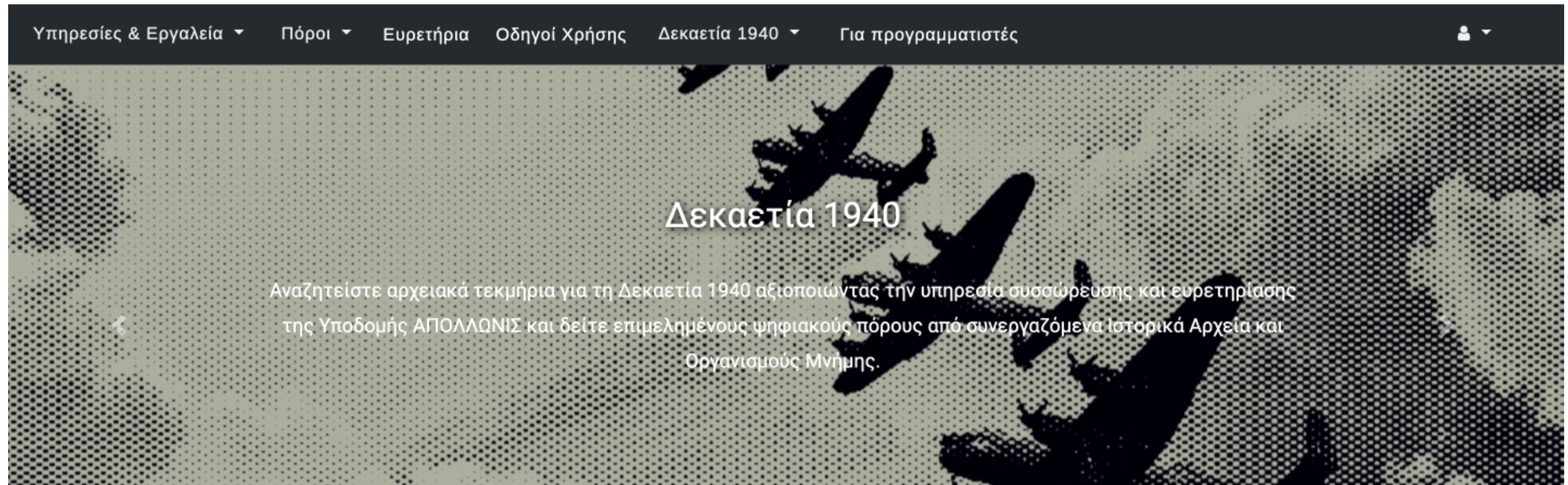
Άλλες προκλήσεις

Καθαρεύουσα, συντομογραφίες, αρκτικόλεξα, κεφαλαιογράμματα γραφή κ.λπ.

Παραδείγματα

Δεκαετία 1940

https://services.apollonis-infrastructure.gr/forties_thema.html



Πρόσωπο

1. Α. Τζήμας (Β. Σαμαρινιώτης ή Σαμαριναίος) Μέλος Στρατιωτικής Διοίκησης Γενικού Στρατηγείου Ε.Λ.Α.Σ.
2. Α. Τζήμας (Β. Σαμαρινιώτης ή Σαμαριναίος) Μέλος Στρατιωτικής Διοίκησης Γενικού Στρατηγείου Ε.Λ.Α.Σ. και Μόνιμος Αντιπρόσωπος Γενικού Στρατηγείου Ε.Λ.Α.Σ. στην Ανώτατη Διοίκηση του Λαϊκού Απελευθερωτικού Στρατού Γιουγκοσλαβίας
3. Τζήμας
4. Τζήμας Ανδρέας (Σαμαρινιώτης)
5. Σαμαριναίος
6. ΤΖΗΜΑΣ



Τόπος

Αθήνα

ΑΘΗΝΑ

Αθήναι

Αθηνών

Αθηναίων

Ακρόπολη

Χρόνος

06/25/44

25-6-1944

25 Ιουνίου 1944

Καλοκαίρι 1944

1944

Σφάλματα

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ **Ι**ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ **ν**

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ **Ω**

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗ **Ι**ΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ Ε **Λ**ΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔ **Α** ΚΑΙ ΣΧΗΜΑΤΙΣΜΟΙ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

ΜΟΝΑΔΕΣ ΚΑΙ ΣΧΗΜΑΤ **Α** ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΣΤΡΑΤΟΥ

Γιατί καθαρισμός;

14.724

Προ καθαρισμού

Ανακτώνται 14.724 αποτελέσματα για Μονάδες και Σχηματισμούς του Ελληνικού Στρατού

17.168

Μετά τον καθαρισμό

Επιστρέφονται 17.1648 αποτελέσματα για Μονάδες και Σχηματισμούς του Ελληνικού Στρατού

Τα βήματα [1/2]

Επιλογή εργαλείου

Επιλέξτε το κατάλληλο εργαλείο ανάλογα με τις ανάγκες σας και το είδος του υλικού σας.

Εισαγωγή δεδομένων

Εξετάστε τους μορφότευπους που υποστηρίζονται από το εργαλείο και εισάγαγετε τα δεδομένα. Επιβεβαιώστε ότι δεν υπήρξαν σφάλματα (π.χ. κωδικοποίησης) κατά την εισαγωγή.

Κανόνες καθαρισμού

Διατυπώστε τους κανόνες καθαρισμού με βάση το υλικό και τους στόχους σας, μετά από μια ενδελεχή επισκόπηση των δεδομένων σας και της δομής τους.

Οριζόντιος καθαρισμός

Επιλέξτε ενέργειες που μπορείτε να εφαρμόσετε οριζόντια σε όλα τα δεδομένα σας (π.χ. τριμάρισμα κενών στην αρχή ή/και το τέλος των τιμών).

Τα βήματα [1/2]

Καθαρισμός ανά πεδίο

Εφαρμόστε καθαρισμό ανά πεδίο με βάση τους κανόνες που διατυπώσατε, τροποποιώντας, εάν χρειάζεται, τους τύπους ή των αριθμό των πεδίων και εφαρμόζοντας κατάλληλες ομαδοποιήσεις και φίλτρα.

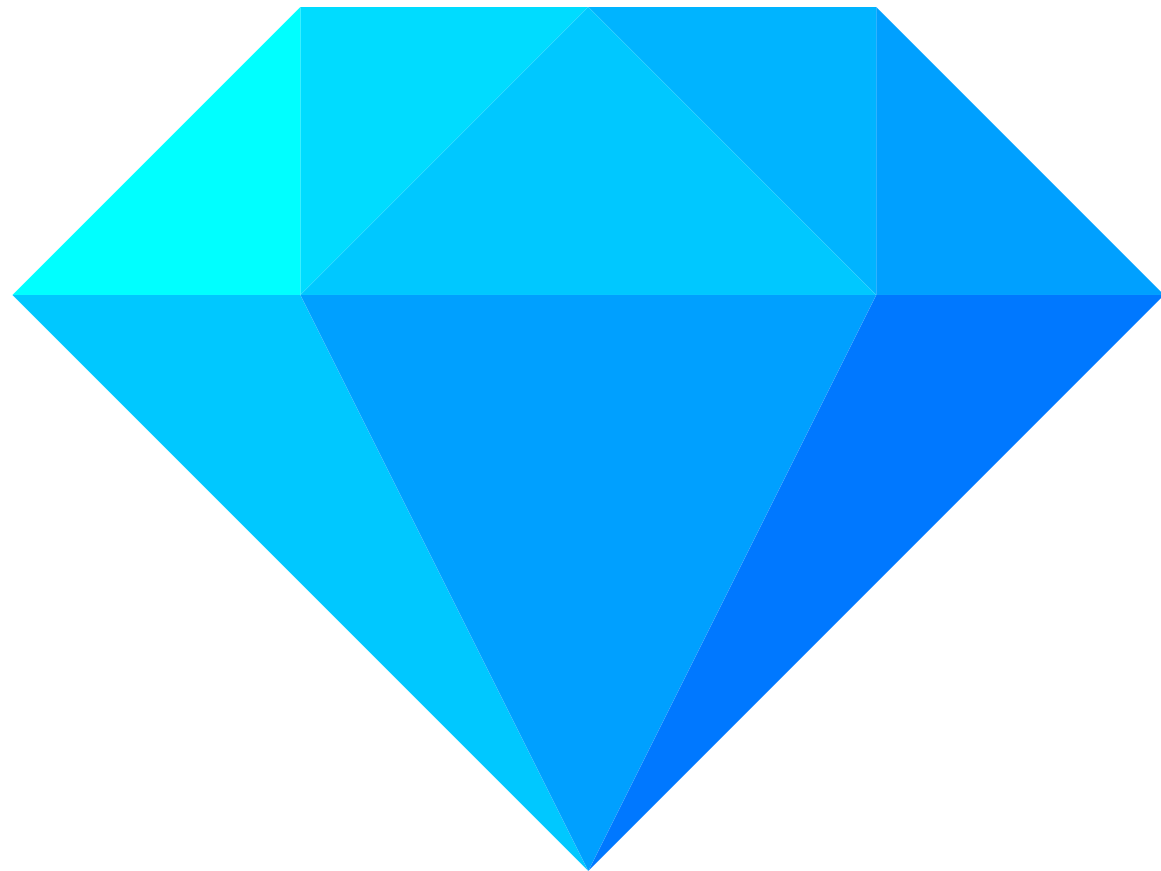
Εξαγωγή αποτελεσμάτων

Εξαγάγετε τα καθαρισμένα δεδομένα σας στον επιθυμητό μορφότυπο.

A pair of blue-handled scissors and a box of blue pencils are positioned in the top right corner of the image. The scissors are open, and the pencils are sharpened and ready for use. The background is a solid light blue color.

Δείτε τα βήματα εδώ

<https://services.apollonis-infrastructure.gr/scenarios.html>



OpenRefine

<https://openrefine.org/>