Δουλεύοντας με messy data – cleaning - filtering

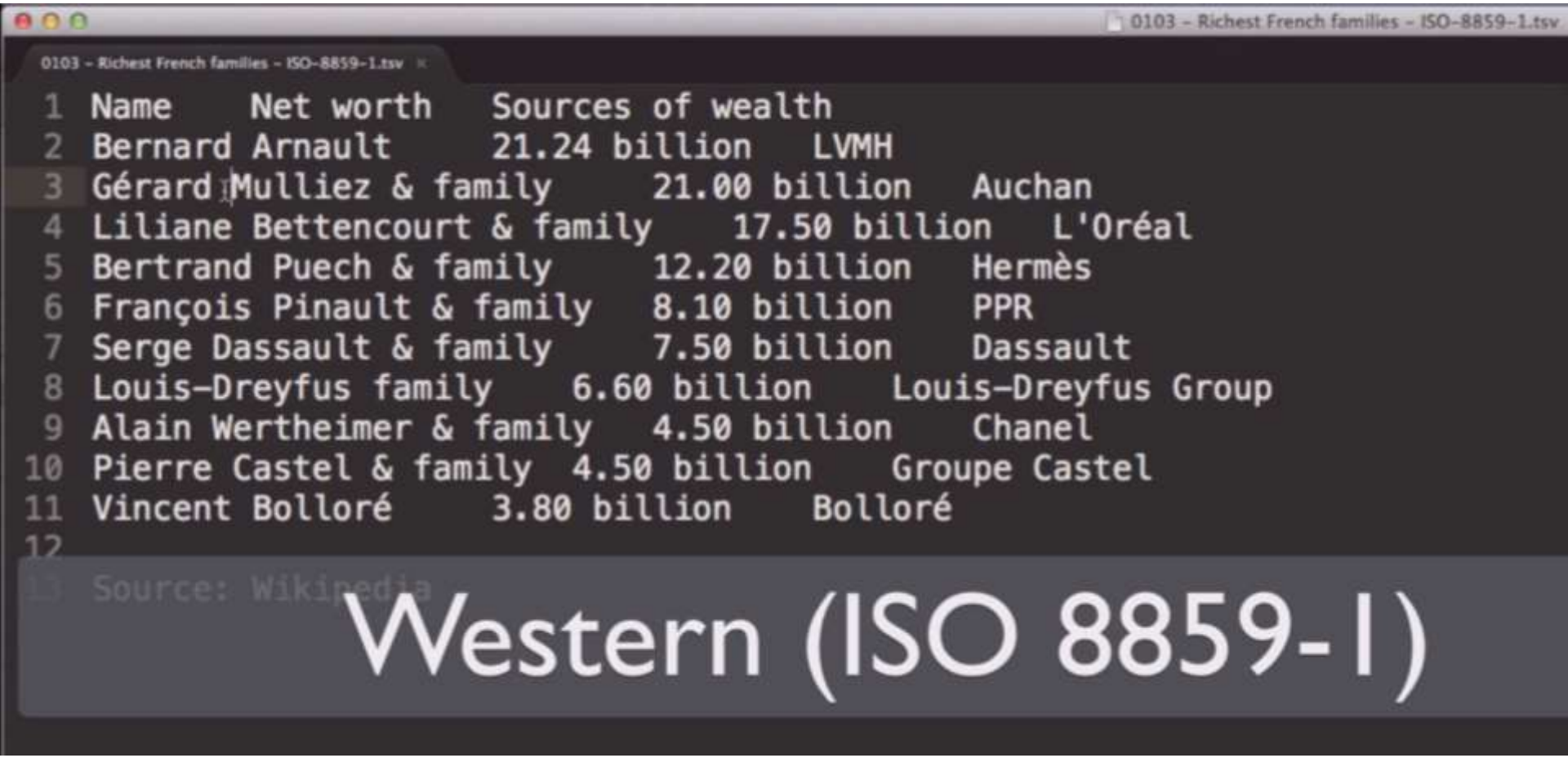
Ανδρέας Βέγλης – καθηγητής

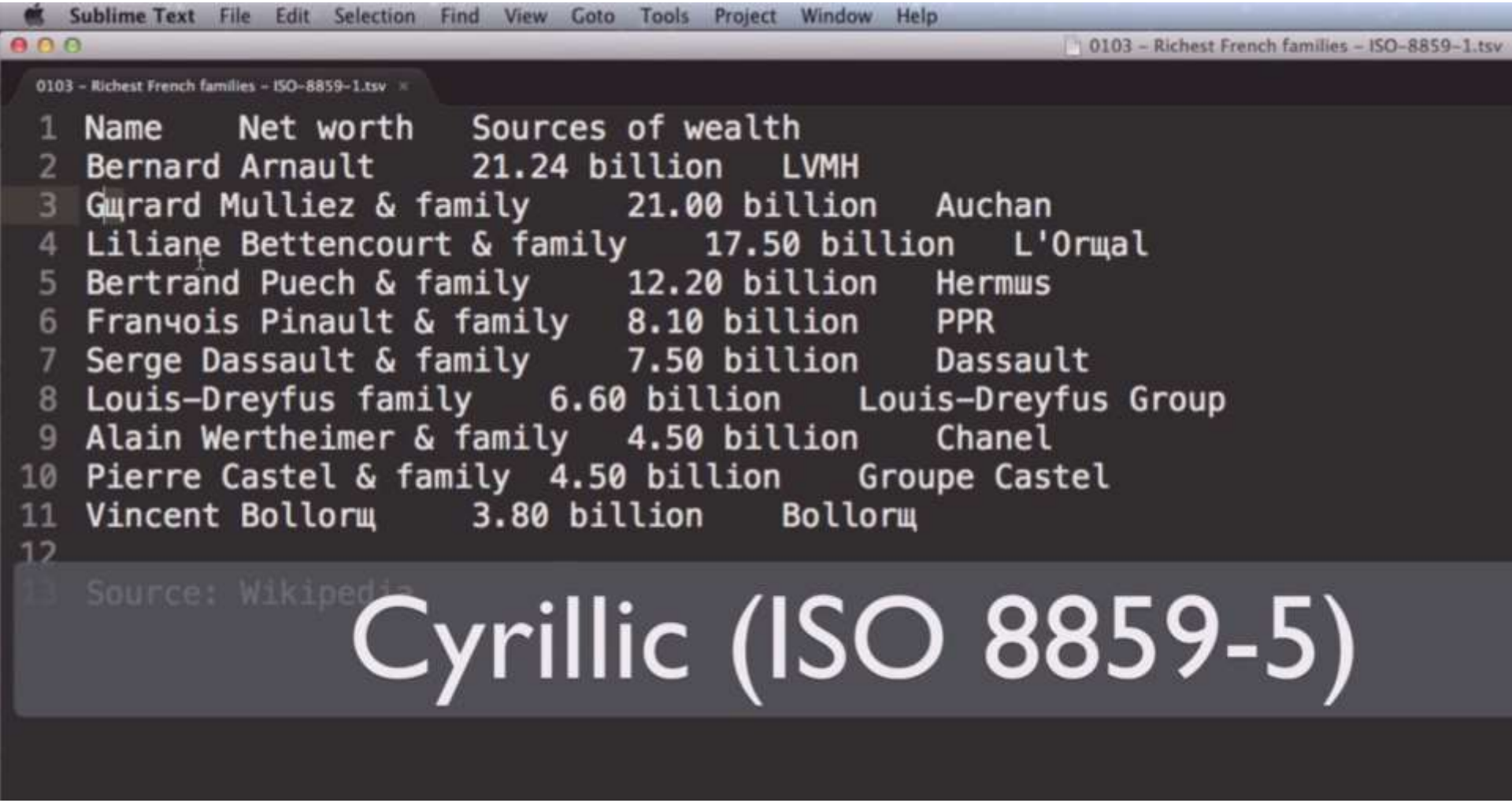
Περιεχόμενο δεδομένων

- Σπάνια τα δεδομένα που συλλέγουμε μπορούν να χρησιμοποιηθούν άμεσα.
- Συνήθη προβλήματα
  - Κωδικοποίηση χαρακτήρων
  - Δεκαδικό διαχωριστικό
  - Ορθογραφικά λάθη

Κωδικοποίηση χαρακτήρων

- Μη σωστή κωδικοποίηση → δεν είναι δυνατή η ανάγνωση των δεδομένων
- Ορίζοντας την κωδικοποίηση χαρακτήρων:
  - Υπολογιστές αναπαριστούν τα πάντα με μηδενικά και μονάδες (αριθμούς και γράμματα).
  - Χαρακτήρας → 1 byte με τιμή από το 1 έως το 255 (00000001- 11111111).
  - Κάθε γράμμα της αλφαβήτου και τα άλλα σύμβολα αναπαριστούνται με ένα τέτοιο byte.
  - a → 97, b → 98, c → 99 έως το 127 είναι τα αγγλικά
- Διαφορές στην κωδικοποίηση των χαρακτήρων ανάμεσα σε διαφορετικές χώρες (από το 128 έως το 255):
  - Στα Γαλλικά το character code 201 είναι το É
  - Ενώ στα Ρωσικά ο character code 201 το Щ


0103 - Richest French families - ISO-8859-1.tsv
0103 - Richest French families - ISO-8859-1.tsv
1 Name Net worth Sources of wealth
2 Bernard Arnault 21.24 billion LVMH
3 Gérard Mulliez & family 21.00 billion Auchan
4 Liliane Bettencourt & family 17.50 billion L'Oréal
5 Bertrand Puech & family 12.20 billion Hermès
6 François Pinault & family 8.10 billion PPR
7 Serge Dassault & family 7.50 billion Dassault
8 Louis-Dreyfus family 6.60 billion Louis-Dreyfus Group
9 Alain Wertheimer & family 4.50 billion Chanel
10 Pierre Castel & family 4.50 billion Groupe Castel
11 Vincent Bolloré 3.80 billion Bolloré
12
13 Source: Wikipedia
Western (ISO 8859-1)


Sublime Text File Edit Selection Find View Goto Tools Project Window Help
0103 - Richest French families - ISO-8859-1.tsv
0103 - Richest French families - ISO-8859-1.tsv
1 Name Net worth Sources of wealth
2 Bernard Arnault 21.24 billion LVMH
3 Gщrard Mulliez & family 21.00 billion Auchan
4 Liliane Bettencourt & family 17.50 billion L'Orщal
5 Bertrand Puech & family 12.20 billion Hermшs
6 Franчois Pinault & family 8.10 billion PPR
7 Serge Dassault & family 7.50 billion Dassault
8 Louis-Dreyfus family 6.60 billion Louis-Dreyfus Group
9 Alain Wertheimer & family 4.50 billion Chanel
10 Pierre Castel & family 4.50 billion Groupe Castel
11 Vincent Bollorщ 3.80 billion Bollorщ
12
13 Source: Wikipedia
Cyrillic (ISO 8859-5)

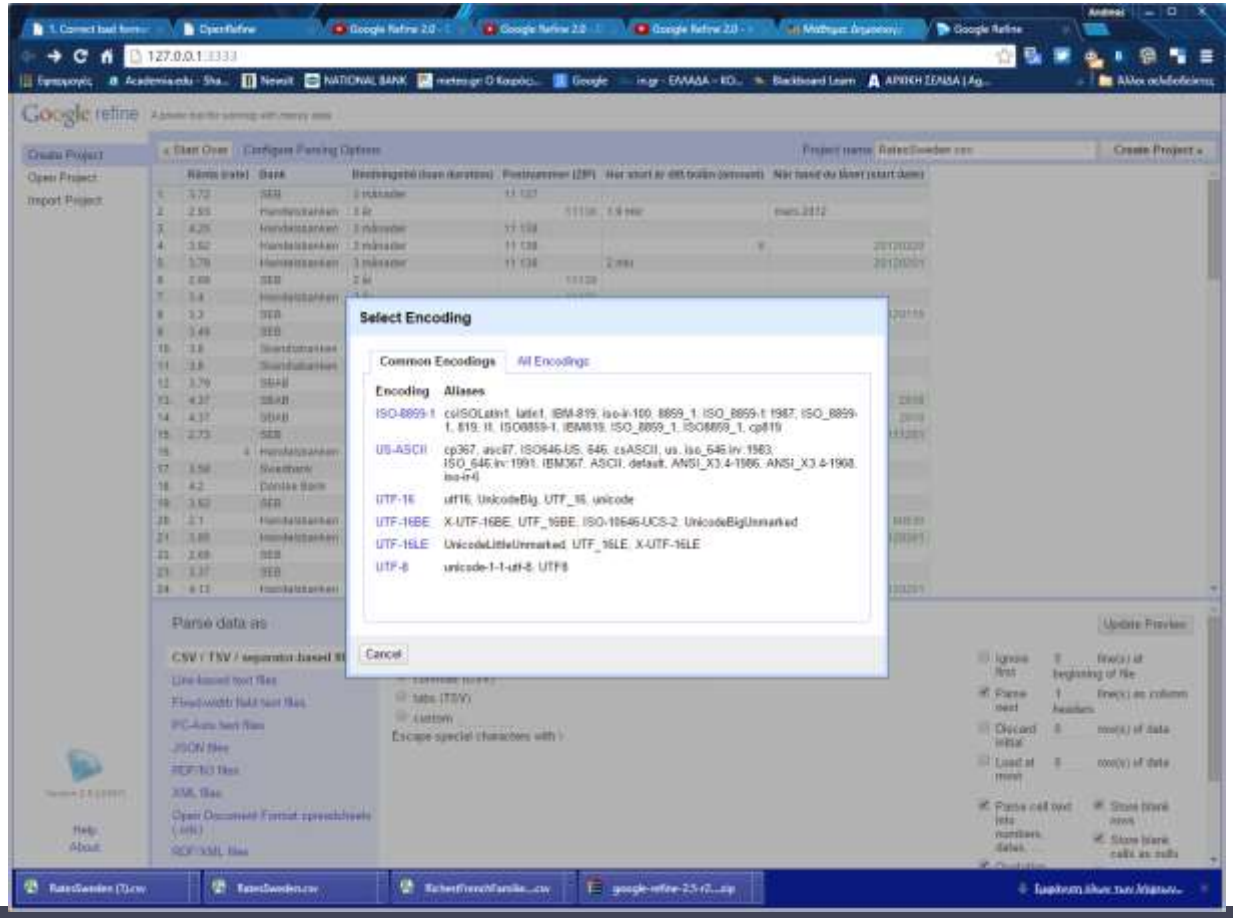
Εύρεση της σωστής κωδικοποίησης

- UTF-8 η στάνταρ κωδικοποίηση που χρησιμοποιείται σήμερα.
- Χρήση του εργαλείου ανοικτού κώδικα Open Refine.

Open Refine

- Open source εργαλείο
- http://openrefine.org
- Portable εφαρμογή (δεν απαιτεί εγκατάσταση)
- Προσφέρει πολλές δυνατότητες επίλυσης προβλημάτων messy data

Αλλαγή κωδικοποιήσεων άμεσα



Δεκαδικό διαχωριστικό

- Άλλες χώρες χρησιμοποιούν το , και άλλες την .
- Χρήση function μετατροπής στο Open Refine:
  - Replace(value, “,”,”.”)
- Δήλωση αριθμών στο Refine

Εισαγωγή αρχείου .csv με κωδικοποίηση UTF-8 στο Excel

- Δεδομένα → Λήψη εξωτερικών δεδομένων από κείμενο → οδηγός εισαγωγής (καθορισμός παραμέτρων)

Διόρθωση ορθογραφικών λαθών

- Αποτέλεσμα:
  - Καθαρισμός των δεδομένων
  - Δυνατότητα για συγκρίσεις

Παράδειγμα Data project

- Slate gun data (Slate)
- Periscopic Gun deaths in the USA data

Μη σωστές τιμής και διπλοεγγραφές

- Πως εντοπίζουμε μη σωστές τιμές στα δεδομένα
- Πως διορθώνουμε τις μη σωστές τιμές καθώς και τις διπλοεγγραφές.

Μη σωστές τιμές

- Πρόβλημα στη διαδικασία εισαγωγής δεδομένων. Ορθογραφικά λάθη.
- Παράδειγμα:
  - Mapped: Every Protest on the Planet Since 1979

Διπλοεγγραφές

- Παράδειγμα, αριθμός μεταλλίων σε Ολυμπιακούς αγώνες.
- Assessment of Italian schools at risk in case of an earthquake

Προχωρημένες τεχνικές καθαρισμού δεδομένων

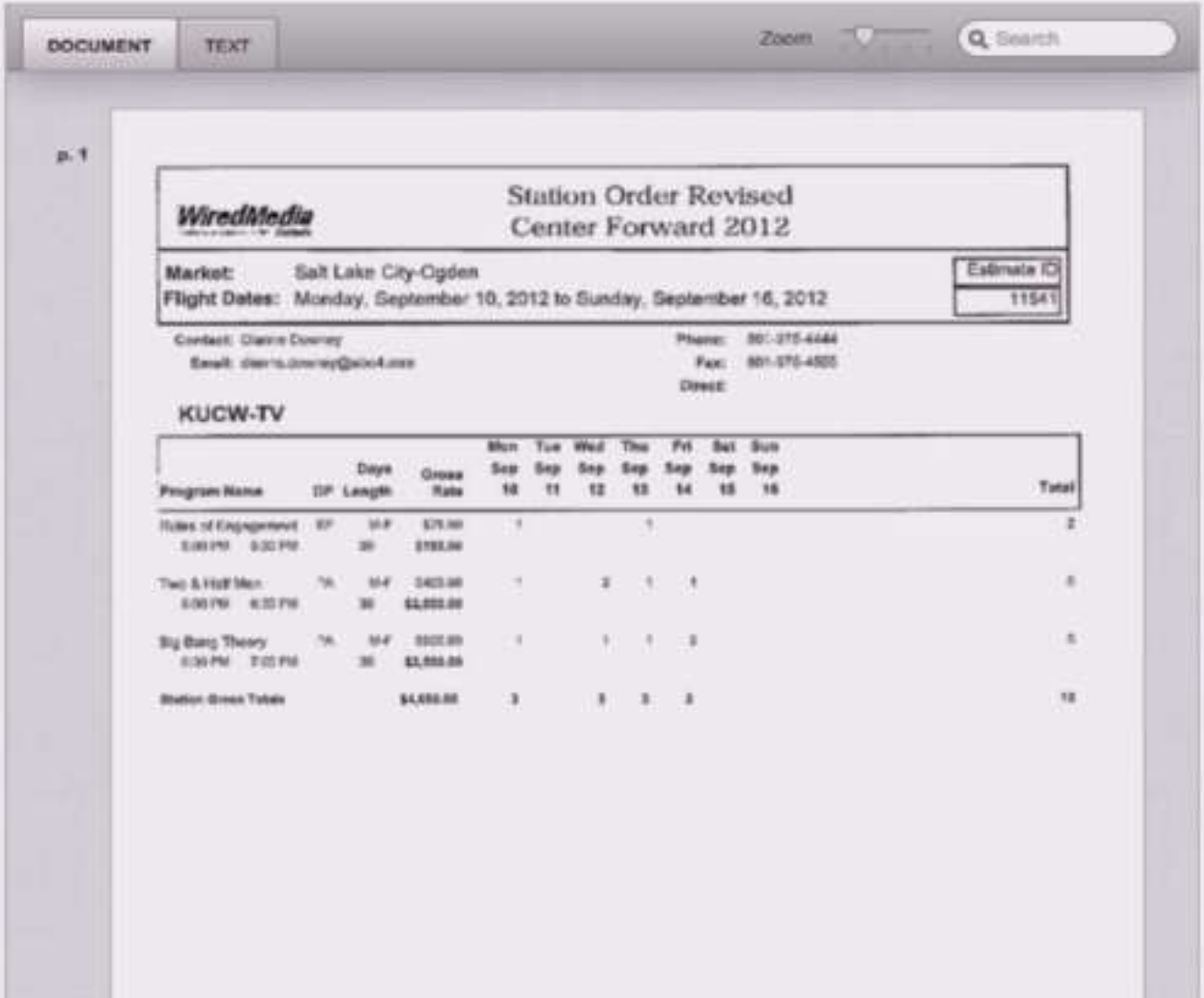
- Χρήστη τεχνικών πληθοπορισμού για τον καθαρισμό δεδομένων.
- Σωστή οργάνωση των δεδομένων σε μορφή (format) βάσης δεδομένων

Παραδείγματα χρήσης πληθοπορισμού για καθαρισμό δεδομένων

- Free the Files: Help ProPublica Unlock Political Ad Spending (καλή εφαρμογή)
- Διαμοιρασμός δεδομένων προς καθαρισμό μέσω google spreasheet (μη καλή εφαρμογή).

Καλή εφαρμογή

Tell Us About This Document



1. Who Bought It?

HELP

Was **CENTER FORWARD** the candidate/committee who bought these ads?

Yes | No

2. What Agency?

HELP

What is the name of the advertising agency that bought these ads?

3. Contract Number?

HELP

What is the contract number on this ad buy? (Sometimes labeled "Rep. Order")

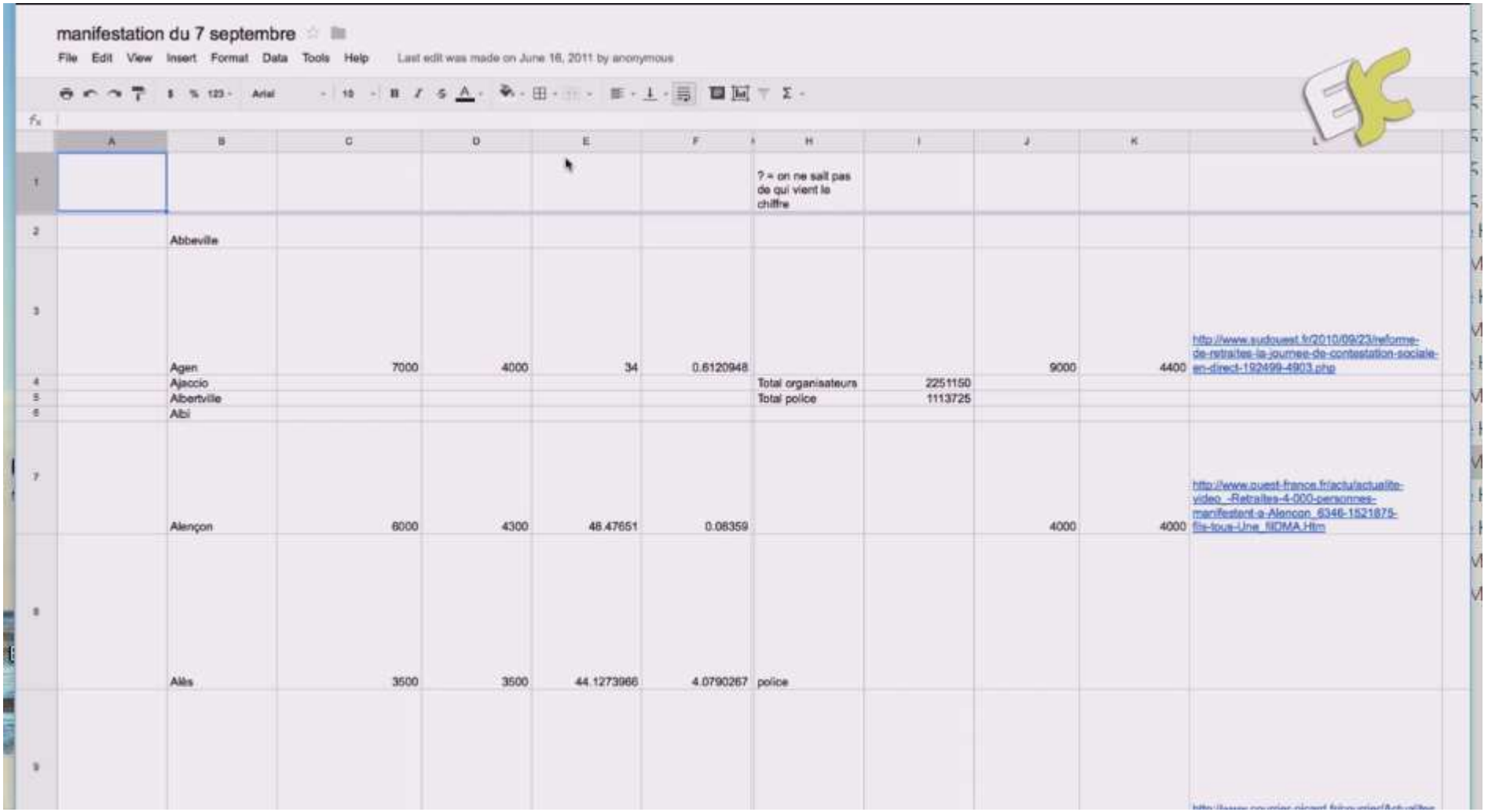
4. How Much?

HELP

What was the *gross* total cost? (Sometimes labeled "Report Totals")

Is there something else notable about this file?

Κακή εφαρμογή

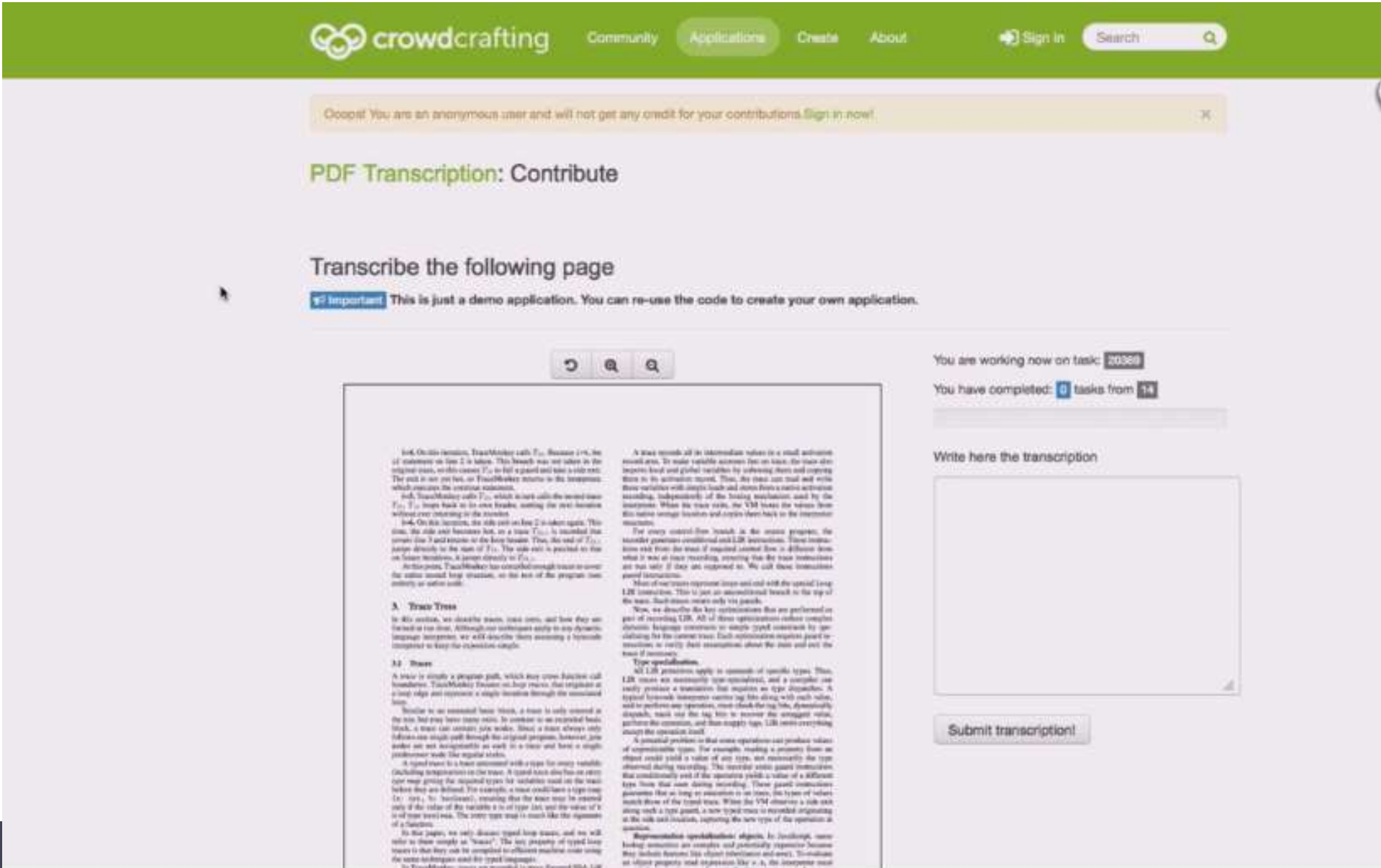


	A	B	C	D	E	F	H	I	J	K	L
1							? = on ne sait pas de qui vient le chiffre				
2		Abbeville									
3											
4		Agen	7000	4000	34	0.6120948			9000	4400	http://www.sudouest.fr/2010/09/23/reforme-de-retraites-la-journee-de-contestation-sociale-en-direct-192499-4903.php
4		Ajaccio					Total organisateurs	2251150			
5		Albertville					Total police	1113725			
6		Albi									
7											
7		Alençon	6000	4300	48.47651	0.08359			4000	4000	http://www.ouest-france.fr/actu/actualite-video_-Retraites-4-000-personnes-manifestent-a-Alencon_6346-1521875-fils-tous-Une_fiDMA.Htm
8											
8		Alès	3500	3500	44.1273966	4.0790267	police				
9											

Εργαλεία για χρήση πληθοπορισμού για καθαρισμό/διόρθωση δεδομένων

- Crowdcrafting
  - Open Knowledge Foundation.
  - Υποστηρίζει τη δημιουργία μικρών διεργασιών για το κοινό.
  - Επιτρέπει την αντιγραφή και τη προσαρμογή γνωστών διεργασιών
- Amazon Mechanical Turk
  - Παρόμοια λογική
  - Υποστηρίζει την αμοιβή των χρηστών που συμμετέχουν.

Crowdcrafting



crowdcrafting
scifabric

DISCOVER ABOUT SIGN IN CREATE YOUR PROJECT

Our Favorite!

Sound Cloud

If you're into music, this is your project. The demo that also shows you how to create your own sound recognition project.

Author

C'mon!

Find a project that fascinates you, and help researchers analyse their data.

FIND A PROJECT

Create a Project and get other volunteers to help you analyse your data.

CREATE A PROJECT

amazonmechanical turk
Artificial Artificial Intelligence

Your Account | HITs | Qualifications

Already have an account?
Sign in as a Worker | Requester

Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

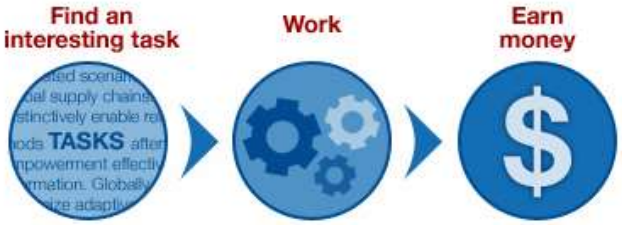
275,829 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Find HITs Now

or learn more about being a **Worker**

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Get Started

An amazon.com company

Συμβουλές

- Καθορίστε ακριβώς την μέθοδο καθαρισμού των δεδομένων.
- Δεν είναι σίγουρο ότι οι συμμετέχοντες χρήστες θα βρούνε τις μη σωστές καταχωρήσεις.
- Μπορεί να χρειαστεί να καθαρίσετε ξανά αυτά που επεξεργάστηκαν οι χρήστες.
- Ο Πληθοπορισμός είναι ενδιαφέρουσα τεχνική αλλά δεν μπορεί να είναι η στάνταρ μέθοδός σας για καθαρισμό δεδομένων.

Η περίπτωση των Μεγάλων Συνόλων Δεδομένων (Big Data Sets)

- Ο τρόπος που τα αντιμετωπίζουμε είναι διαφορετικός από την περίπτωση των μικρών σετ δεδομένων.
- Χωρισμός των δεδομένων σε πολλούς συνδεδεμένους πίνακες.
- Διευκολύνει το καθαρισμό των δεδομένων.