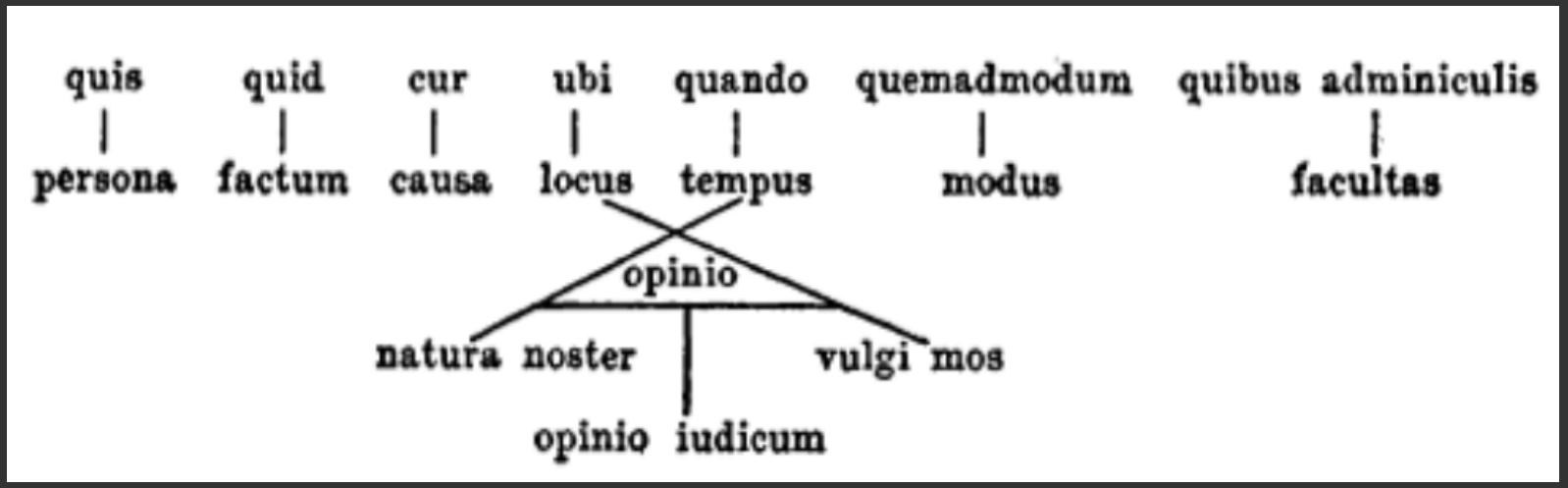
FONDAZIONE
BRUNO KESSLER

Spreadsheets are your friends

(and your data will love them)

Cristian Consonni
Fondazione Bruno Kessler
20 dicembre 2013
School of data, Trento

Struttura della notizia

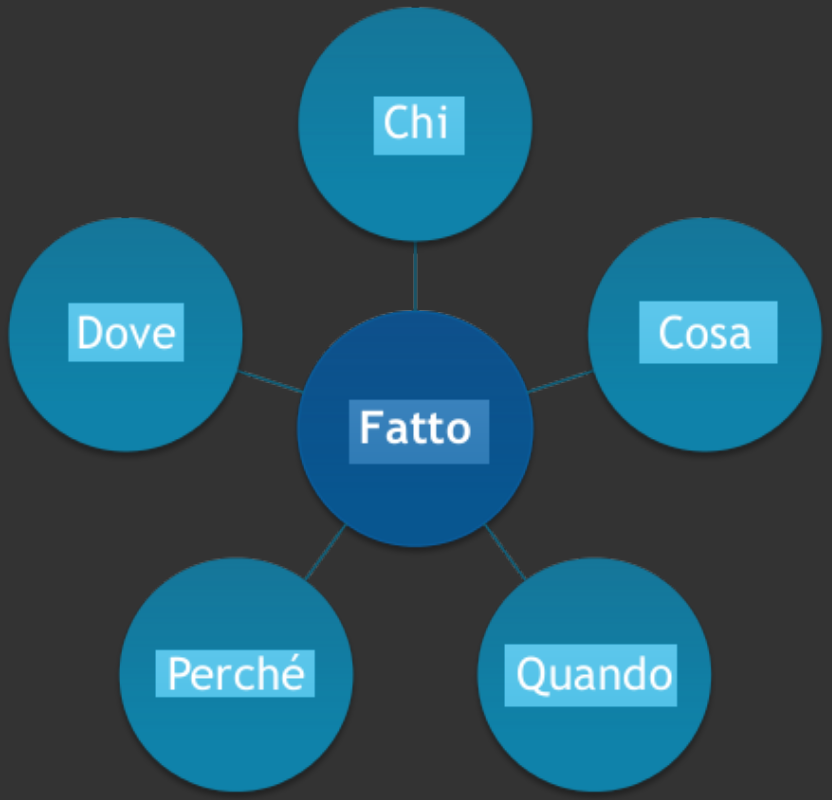


Le 5 “W”:

- Who is it about?
- What happened?
- When did it take place?
- Where did it take place?
- Why did it happen?

Struttura della notizia

- Ogni aspetto di una notizia può essere tradotto in un dato ↔ un dato può essere incorporato in un aspetto della notizia;
- Ogni colonna è una dimensione dei dati;
- I dati devono aiutare a rispondere alle domande precedenti;



«The problem I have is that the worldview that my students have correspond to reality in the world the year their teachers were born»

http://www.gapminder.org/videos/ted-us-state-department/

Perché i dati?

«Software is what the 21st century is made of.

What steel was to the economy of the 20th century.

What steel was to the power of the 20th century

What steel was to the politics of the 20th century, software is now.

It's the crucial building block, the component out of which everything else is made.

And when I speak of everything, else I mean, of course, freedom.»

Tratto da:
"Why Political Liberty Depends on Software Freedom More Than Ever"

Eben Moglen @ 2011 FOSDEM conference in Brussels on Feb 5, 2011

http://www.softwarefreedom.org/events/2011/fosdem/moglen-fosdem-keynote.html

SCHOOL OF DATA

Esercitazione

FONDAZIONE BRUNO KESSLER

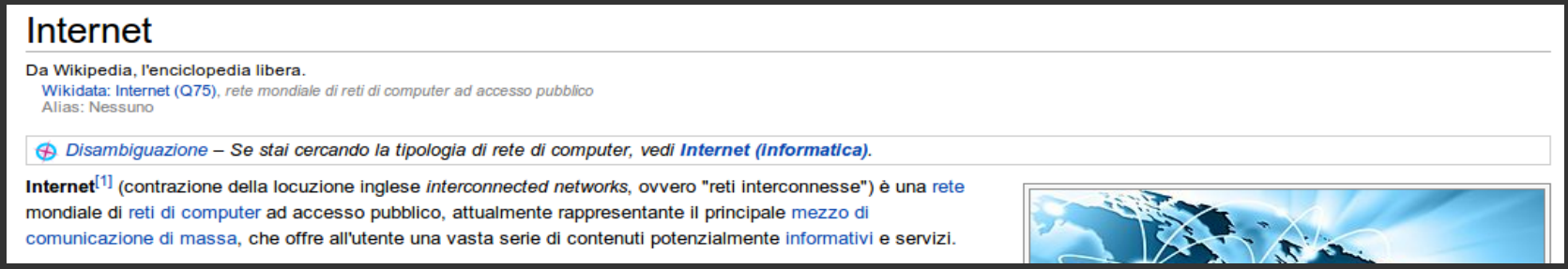
Data pipeline: summary

- Data pipeline I: acquisition
- Data pipeline II: cleaning
- Data pipeline III: analysis
- Data pipeline IV: visualizing

Data Acquisition: forma dei dati

- Human-readable

Dati leggibili facilmente da un umano, per esempio, una pagina di Wikipedia.



Internet

Da Wikipedia, l'enciclopedia libera.
Wikidata: Internet (Q75), *rete mondiale di reti di computer ad accesso pubblico*
Alias: Nessuno

Disambiguazione – Se stai cercando la tipologia di rete di computer, vedi ***Internet (Informatica)****.*

Internet[1] (contrazione della locuzione inglese *interconnected networks*, ovvero "reti interconnesse") è una rete mondiale di reti di computer ad accesso pubblico, attualmente rappresentante il principale mezzo di comunicazione di massa, che offre all'utente una vasta serie di contenuti potenzialmente informativi e servizi.

- Machine-readable

«Formats that are machine readable are ones which are able to have their data extracted by computer programs easily. […] Common machine-readable file formats are CSV files.»
da http://schoolofdata.org/handbook/appendix/glossary/#term-machine-readable

Data acquisition: obiettivo finale

La data acquisition consiste nell'ottenere dei dati in formato *machine-readable*

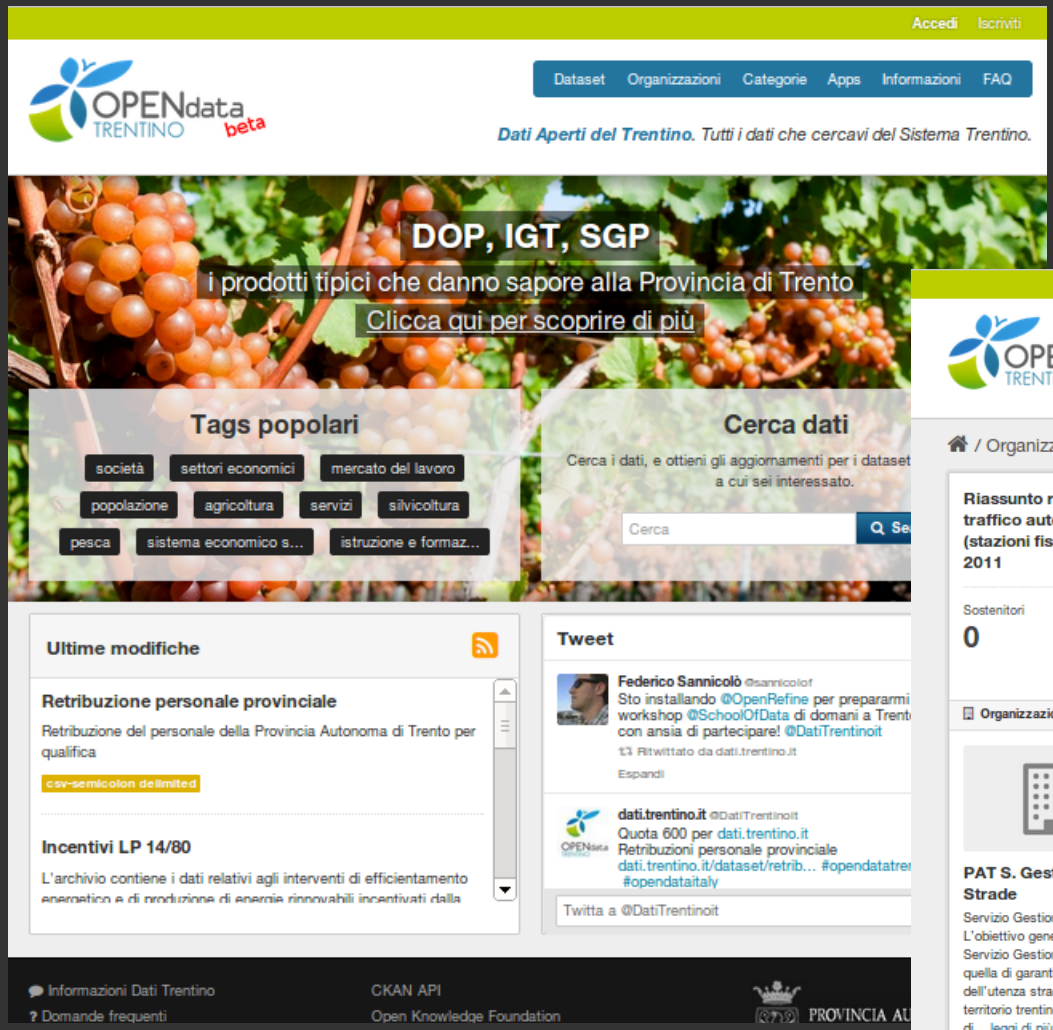
Metodi:

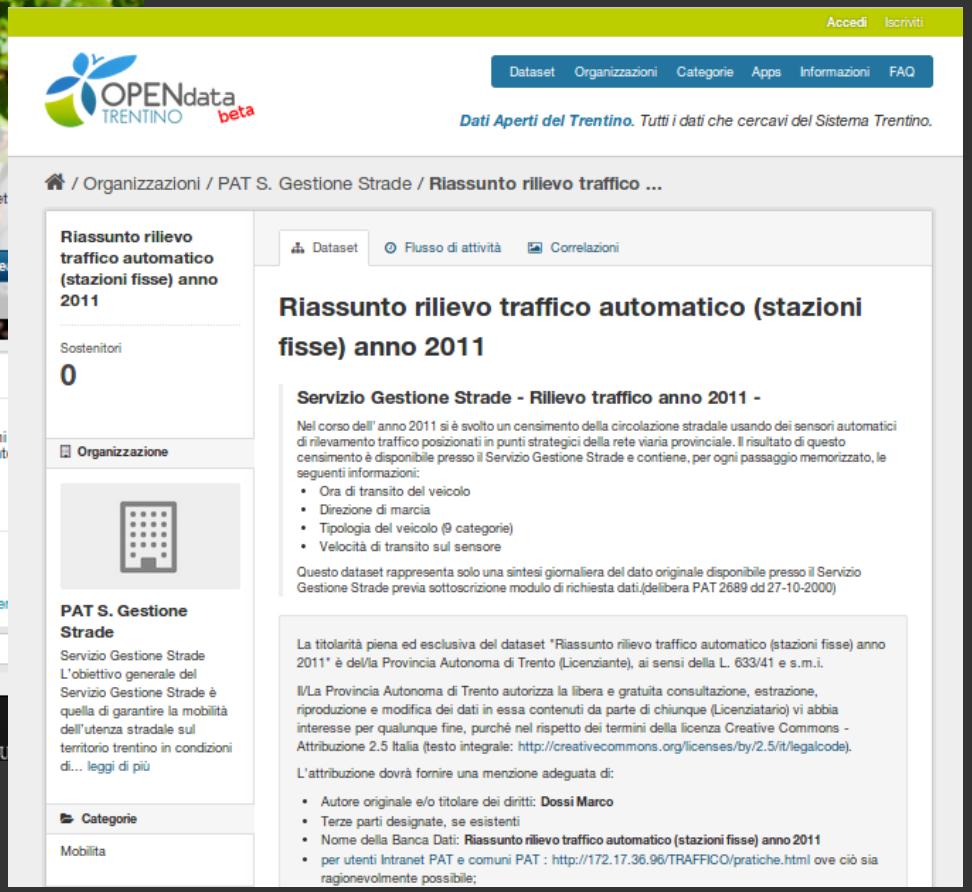
- Scaricare dataset da portali open-data (facile)
- Scraping di pagine web (medio)
- Scraping di PDF (difficile)

Acquisition: good questions

- Chi ha prodotto i dati? Un ente pubblico? Un azienda? (affidabilità)
- Come sono stati prodotti i dati? Il processo di raccolta dati è documentato?
- È possibile ottenere gli stessi dati (o almeno dati simili) in altri modi? È possibile confrontare dati di dettaglio con dati aggregati?

Datasets


Accedi Iscriviti
OPENdata TRENTINO beta
Dataset Organizzazioni Categorie Apps Informazioni FAQ
Dati Aperti del Trentino. Tutti i dati che cercavi del Sistema Trentino.
DOP, IGT, SGP
i prodotti tipici che danno sapore alla Provincia di Trento
Clicca qui per scoprire di più
Tags popolari
società
settori economici
mercato del lavoro
popolazione
agricoltura
servizi
silvicoltura
pesca
sistema economico s...
istruzione e formaz...
Cerca dati
Cerca i dati, e ottieni gli aggiornamenti per i dataset a cui sei interessato.
Cerca
Ultime modifiche
Retribuzione personale provinciale
Retribuzione del personale della Provincia Autonoma di Trento per qualifica
csv-semicolon delimited
Incentivi LP 14/80
L'archivio contiene i dati relativi agli interventi di efficientamento
Tweet
Federico Sannicolò @sannicolof
Sto installando @OpenRefine per prepararmi workshop @SchoolOfData di domani a Trent con ansia di partecipare! @DatiTrentinoit
Ritwittato da dati.trentino.it
Espandi
dati.trentino.it @DatiTrentinoit
Quota 600 per dati.trentino.it
Retribuzioni personale provinciale
dati.trentino.it/dataset/retrib... #opendatatre #opendataitaly
Twitta a @DatiTrentinoit
Informazioni Dati Trentino
Domande frequenti
CKAN API
Open Knowledge Foundation
PROVINCIA AU


Accedi Iscriviti
OPENdata TRENTINO beta
Dataset Organizzazioni Categorie Apps Informazioni FAQ
Dati Aperti del Trentino. Tutti i dati che cercavi del Sistema Trentino.
/ Organizzazioni / PAT S. Gestione Strade / Riassunto rilievo traffico ...
Riassunto rilievo traffico automatico (stazioni fisse) anno 2011
Sostenitori
0
Organizzazione
PAT S. Gestione Strade
Servizio Gestione Strade L'obiettivo generale del Servizio Gestione Strade è quella di garantire la mobilità dell'utenza stradale sul territorio trentino in condizioni di... leggi di più
Categorie
Mobilità
Dataset
Flusso di attività
Correlazioni
Riassunto rilievo traffico automatico (stazioni fisse) anno 2011
Servizio Gestione Strade - Rilievo traffico anno 2011 -
Nel corso dell' anno 2011 si è svolto un censimento della circolazione stradale usando dei sensori automatici di rilevamento traffico posizionati in punti strategici della rete viaria provinciale. Il risultato di questo censimento è disponibile presso il Servizio Gestione Strade e contiene, per ogni passaggio memorizzato, le seguenti informazioni:
Ora di transito del veicolo
Direzione di marcia
Tipologia del veicolo (9 categorie)
Velocità di transito sul sensore
Questo dataset rappresenta solo una sintesi giornaliera del dato originale disponibile presso il Servizio Gestione Strade previa sottoscrizione modulo di richiesta dati.(delibera PAT 2689 dd 27-10-2000)
La titolarità piena ed esclusiva del dataset "Riassunto rilievo traffico automatico (stazioni fisse) anno 2011" è del/la Provincia Autonoma di Trento (Licenziante), ai sensi della L. 633/41 e s.m.i.
Il/La Provincia Autonoma di Trento autorizza la libera e gratuita consultazione, estrazione, riproduzione e modifica dei dati in essa contenuti da parte di chiunque (Licenziatario) vi abbia interesse per qualunque fine, purché nel rispetto dei termini della licenza Creative Commons - Attribuzione 2.5 Italia (testo integrale: http://creativecommons.org/licenses/by/2.5/it/legalcode).
L'attribuzione dovrà fornire una menzione adeguata di:
Autore originale e/o titolare dei diritti: Dossi Marco
Terze parti designate, se esistenti
Nome della Banca Dati: Riassunto rilievo traffico automatico (stazioni fisse) anno 2011
per utenti Intranet PAT e comuni PAT : http://172.17.36.96/TRAFFICO/pratiche.html ove ciò sia ragionevolmente possibile;

File CSV

CSV (formato testo)

http://dati.trentino.it/it/storage/f/2013-11-11T155543/riassunto_dati_traffico_anno_2011.csv

riassunto_dati_traffico_anno_2011.csv (~/Scrivania/schoolofdata/data) - gedit

File Modifica Visualizza Cerca Strumenti Documenti Aiuto

Apri Salva Annulla

riassunto_dati_traffico_anno_2011.csv

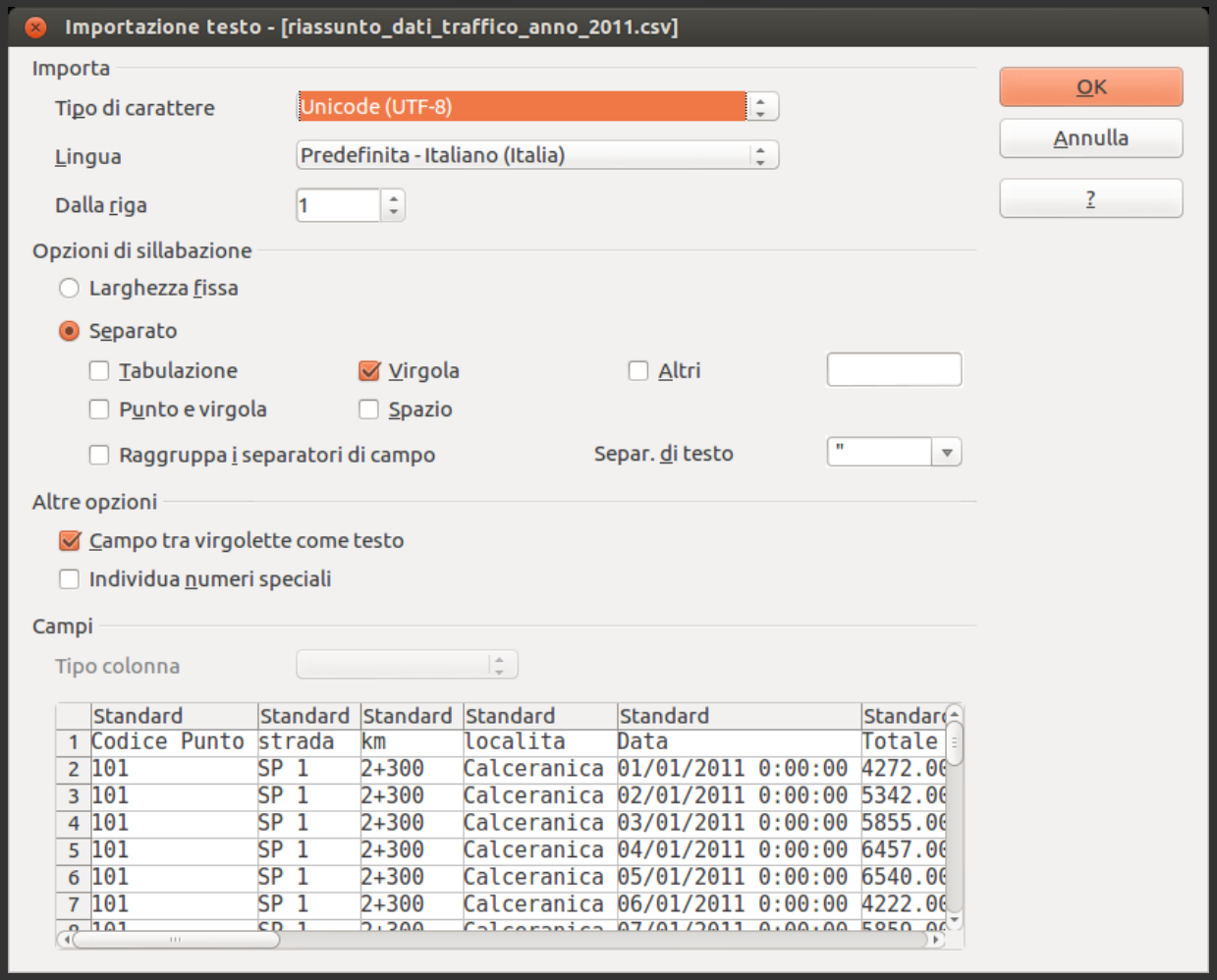
```
1 Codice Punto,strada,km,localita,Data,Totale,Motocicli,Autovetture e monovolumi,Autovetture e monovolumi con
  rimorchio,Furgoni,Autocarro medio (fino a 8_7 m),Autocarro grande (da 8_7 m),Autocarro con rimorchio,Trattore con semirimorchio,Autobus
2 101,SP 1,2+300,Calceranica,01/01/2011 0:00:00,4272.00,44.00,4133.00,1.00,70.00,9.00,2.00,1.00,1.00,11.00
3 101,SP 1,2+300,Calceranica,02/01/2011 0:00:00,5342.00,38.00,5167.00,7.00,93.00,19.00,3.00,1.00,1.00,13.00
4 101,SP 1,2+300,Calceranica,03/01/2011 0:00:00,5855.00,41.00,5488.00,15.00,191.00,62.00,26.00,2.00,9.00,21.00
5 101,SP 1,2+300,Calceranica,04/01/2011 0:00:00,6457.00,53.00,6006.00,21.00,184.00,95.00,53.00,8.00,17.00,20.00
6 101,SP 1,2+300,Calceranica,05/01/2011 0:00:00,6540.00,51.00,6118.00,22.00,209.00,78.00,23.00,8.00,8.00,23.00
7 101,SP 1,2+300,Calceranica,06/01/2011 0:00:00,4222.00,60.00,4025.00,8.00,84.00,22.00,12.00,0.00,1.00,10.00
8 101,SP 1,2+300,Calceranica,07/01/2011 0:00:00,5859.00,49.00,5503.00,17.00,158.00,63.00,38.00,2.00,9.00,20.00
9 101,SP 1,2+300,Calceranica,08/01/2011 0:00:00,5834.00,48.00,5540.00,17.00,141.00,39.00,18.00,0.00,10.00,21.00
10 101,SP 1,2+300,Calceranica,09/01/2011 0:00:00,4422.00,50.00,4262.00,3.00,84.00,8.00,2.00,3.00,3.00,7.00
11 101,SP 1,2+300,Calceranica,10/01/2011 0:00:00,6210.00,43.00,5710.00,6.00,256.00,100.00,42.00,7.00,19.00,27.00
12 101,SP 1,2+300,Calceranica,11/01/2011 0:00:00,6477.00,56.00,5921.00,20.00,252.00,113.00,52.00,12.00,24.00,27.00
13 101,SP 1,2+300,Calceranica,12/01/2011 0:00:00,6495.00,62.00,5899.00,21.00,310.00,113.00,35.00,9.00,22.00,24.00
14 101,SP 1,2+300,Calceranica,13/01/2011 0:00:00,6456.00,56.00,5866.00,20.00,287.00,128.00,36.00,8.00,25.00,30.00
15 101,SP 1,2+300,Calceranica,14/01/2011 0:00:00,7148.00,77.00,6490.00,31.00,306.00,128.00,54.00,7.00,25.00,30.00
16 101,SP 1,2+300,Calceranica,15/01/2011 0:00:00,6475.00,96.00,6107.00,8.00,165.00,54.00,11.00,2.00,11.00,21.00
17 101,SP 1,2+300,Calceranica,16/01/2011 0:00:00,5414.00,75.00,5217.00,6.00,90.00,14.00,3.00,1.00,0.00,8.00
18 101,SP 1,2+300,Calceranica,17/01/2011 0:00:00,6225.00,57.00,5622.00,15.00,292.00,129.00,48.00,8.00,23.00,31.00
19 101,SP 1,2+300,Calceranica,18/01/2011 0:00:00,6597.00,51.00,5874.00,33.00,370.00,138.00,66.00,10.00,24.00,31.00
20 101,SP 1,2+300,Calceranica,19/01/2011 0:00:00,6416.00,73.00,5791.00,24.00,297.00,118.00,30.00,9.00,31.00,23.00
   +300,Calceranica,20  724.00,71.00,605  129.00,53.00,1  0
```

import nel foglio di calcolo

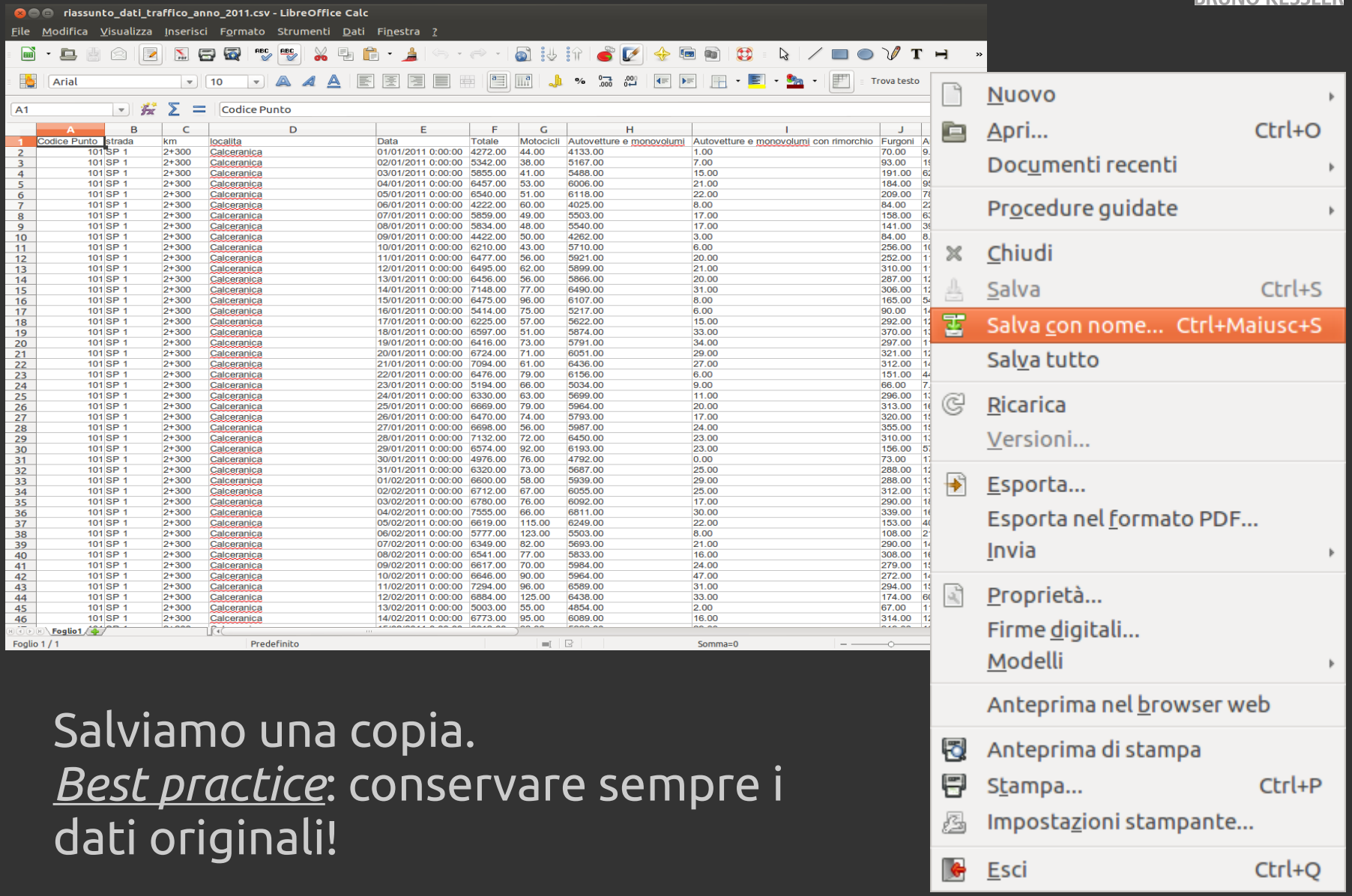
Usiamo LibreOffice:
www.libreoffice.org

import nel foglio di calcolo (II)

Aprire il CSV con LibreOffice Calc: parte la procedura guidata



import nel foglio di calcolo (III)

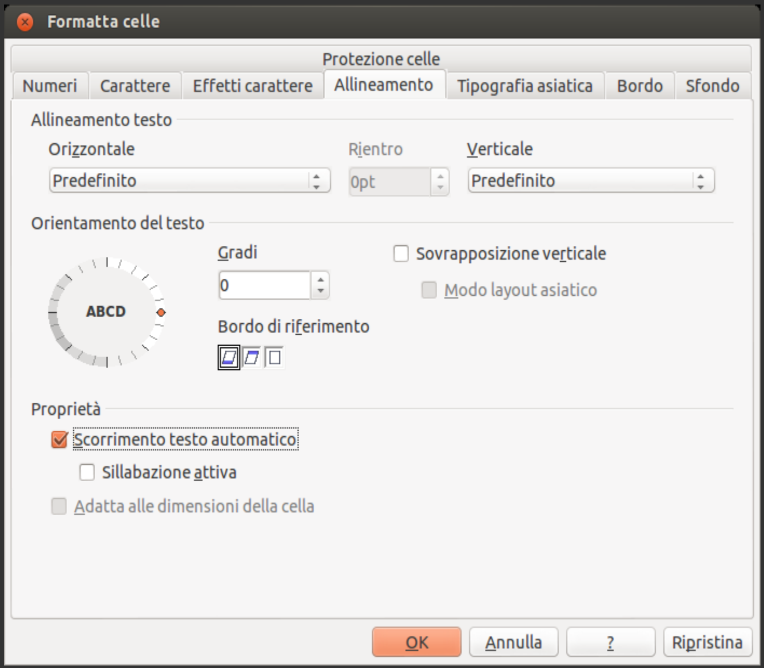


Salviamo una copia.

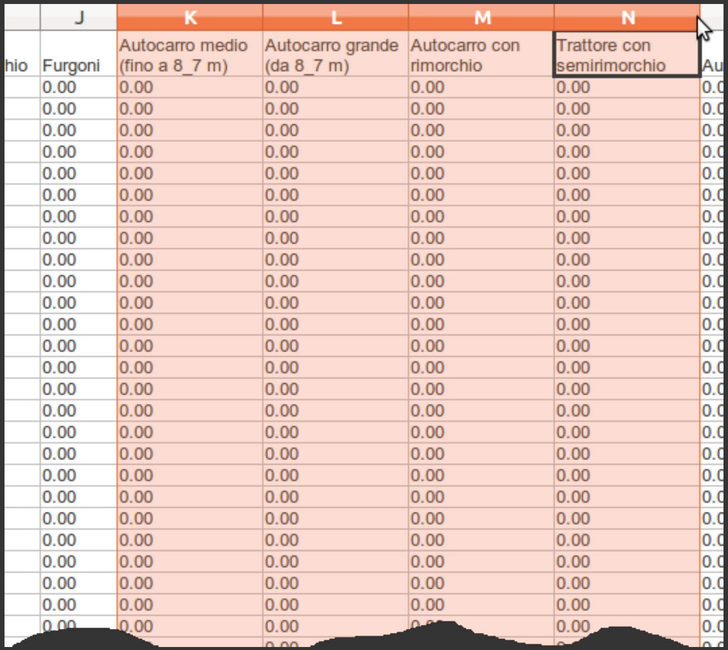
Best practice: conservare sempre i dati originali!

Tricks

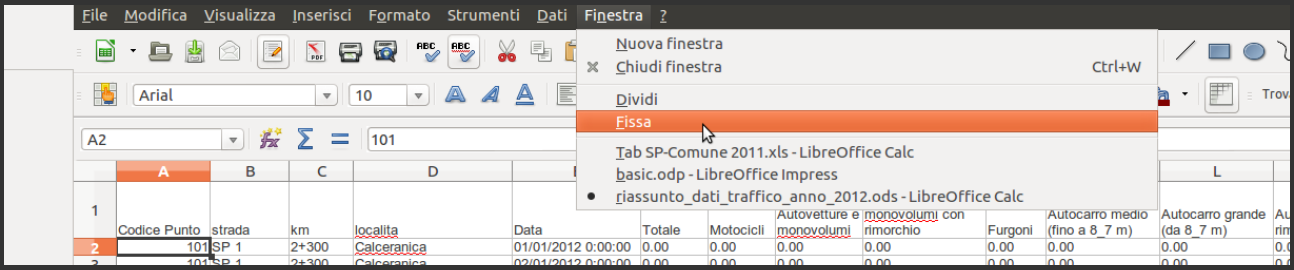
Allineamento celle



Ridimensionare le colonne



Fissare le intestazioni



Data type (I)

Facciamo delle somme

28304		177	S	5/12/2012 0:00:00	1223.00	34.00	1105.00	8.00
28305		177		6/12/2012 0:00:00	2132.00	35.00	2026.00	7.00
28306		17		/12/2012 0:00:00	1243.00	32.00	1110.00	2.00
28307		177		2/2012 0:00:00	1387.00	28.00	1213.00	7.00
28308		177		/2012 0:00:00	1496.00	37.00	1319.00	7.00
28309		177	S	2012 0:00:00	1397.00	28.00	1244.00	8.00
28310		177	S	2012 0:00:00	1468.00	57.00	1278.00	10.0
28311		177	S	/2012 0:00:00	1620.00	50.00	1507.00	7.00
28312		177	SS	2/2012 0:00:00	2187.00	61.00	2083.00	4.00
28313		177	SS	/12/2012 0:00:00	1356.00	34.00	1252.00	7.00
28314		177	SS	5/12/2012 0:00:00	1181.00	31.00	1134.00	0.00
28315		177	SS	/12/2012 0:00:00	1367.00	45.00	1299.00	4.00
28316		177	SS	2/2012 0:00:00	1465.00	32.00	1349.00	5.00
28317		177	SS	2/2012 0:00:00	1718.00	41.00	1559.00	15.0
28318		177	SS	/2012 0:00:00	1804.00	57.00	1677.00	15.0
28319		177	SS	2/2012 0:00:00	2106.00	54.00	2005.00	4.00
28320		177	SS	1/12/2012 0:00:00	1816.00	57.00	1687.00	5.00
28321	TOTALE					0	0	

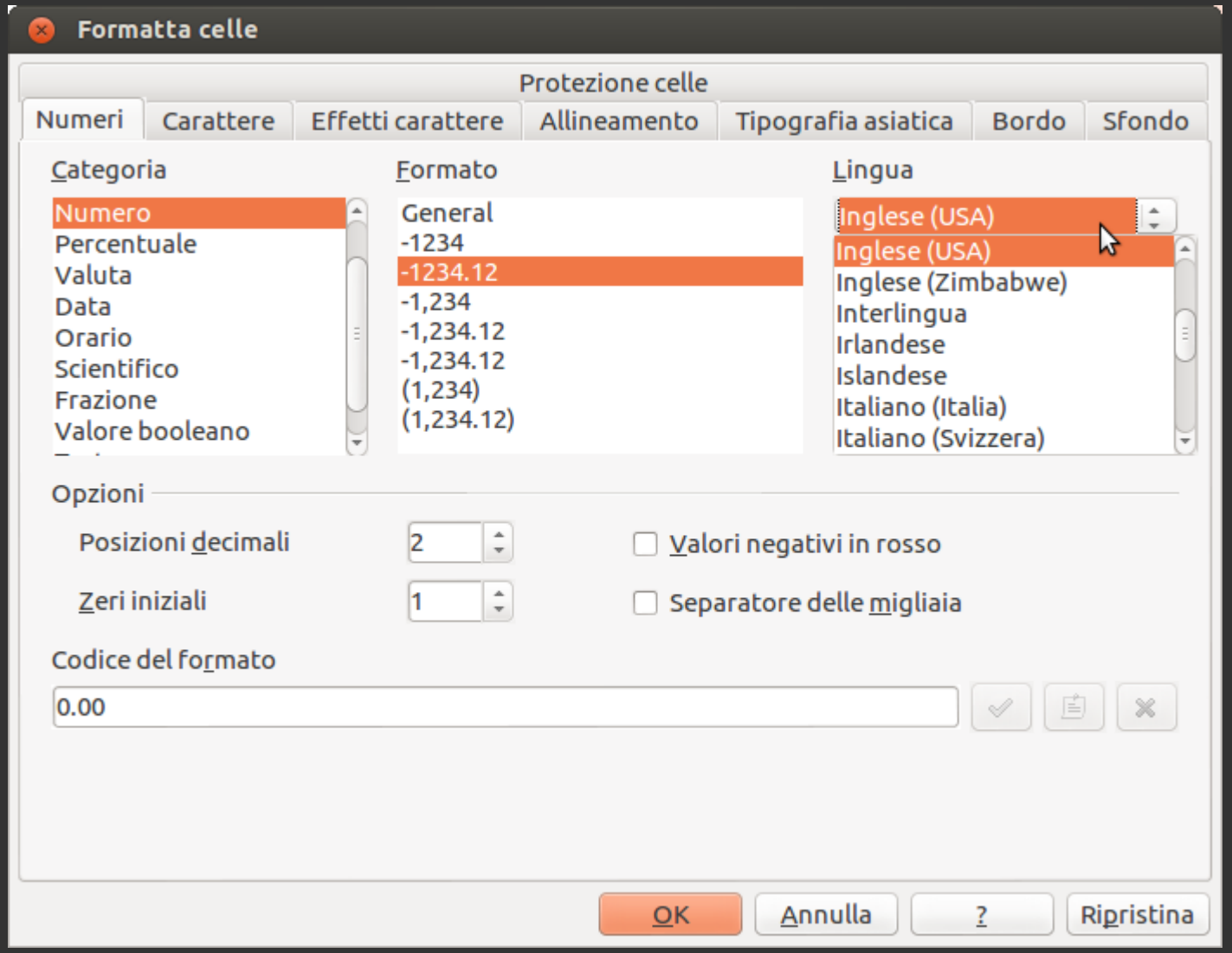
Data type (I)

Facciamo delle somme

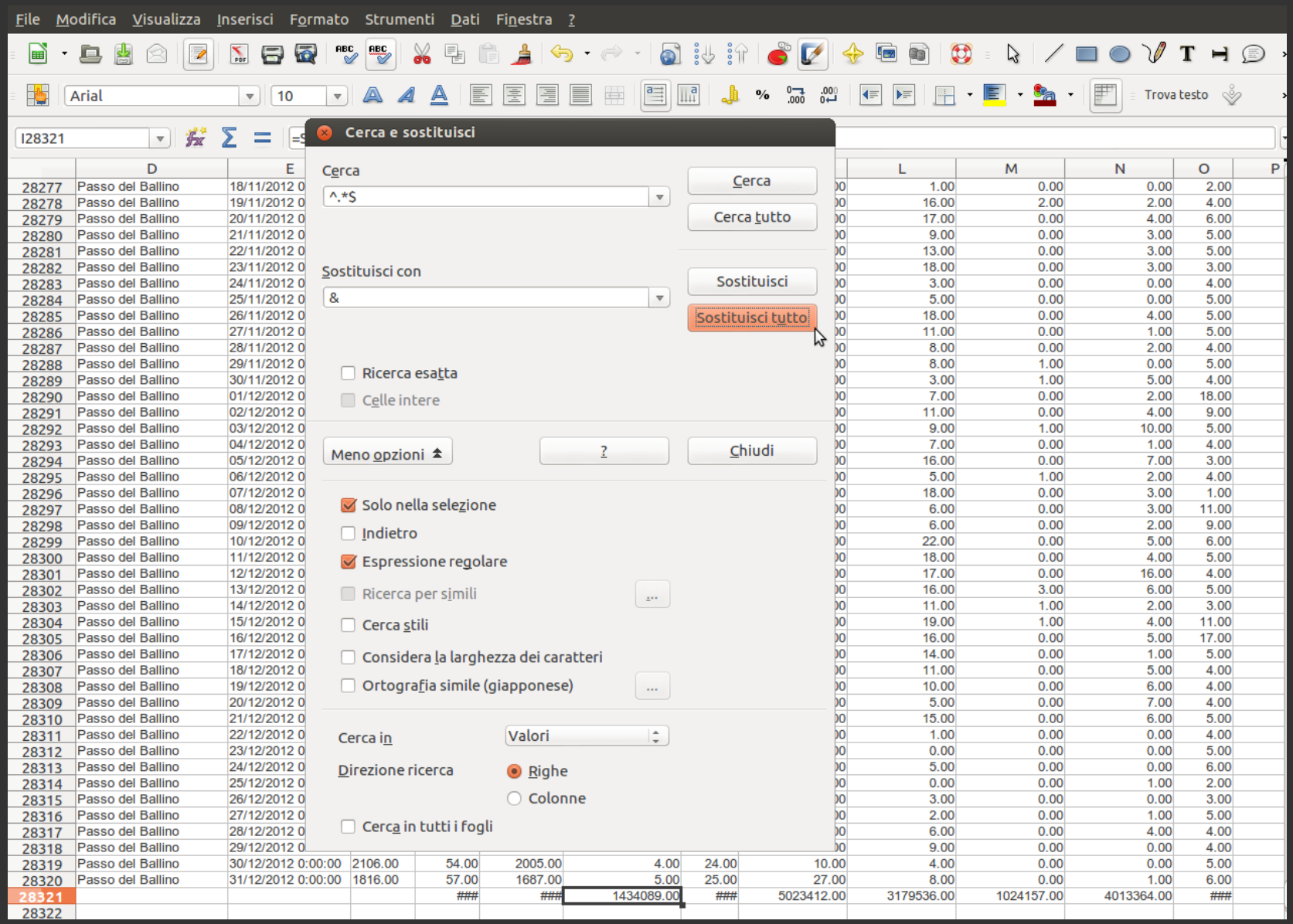
28304		177	S	5/12/2012 0:00:00	1223.00	34.00	1105.00	8.00
28305		177		6/12/2012 0:00:00	2132.00	35.00	2026.00	7.00
28306		177		/12/2012 0:00:00	1243.00	32.00	1110.00	2.00
28307		177		2/2012 0:00:00	1387.00	28.00	1213.00	7.00
28308		177		/2012 0:00:00	1496.00	37.00	1319.00	7.00
28309		177	S	2012 0:00:00	1397.00	28.00	1244.00	8.00
28310		177	S	2012 0:00:00	1468.00	57.00	1278.00	10.0
28311		177	S	/2012 0:00:00	1620.00	50.00	1507.00	7.00
28312		177	SS	2/2012 0:00:00	2187.00	61.00	2083.00	4.00
28313		177	SS	/12/2012 0:00:00	1356.00	34.00	1252.00	7.00
28314		177	SS	5/12/2012 0:00:00	1181.00	31.00	1134.00	0.00
28315		177	SS	/12/2012 0:00:00	1367.00	45.00	1299.00	4.0
28316		177	SS	12/2012 0:00:00	1465.00	32.00	1349.00	.00
28317		177	SS	2/2012 0:00:00	1718.00	41.00	1559.00	15.0
28318		177	SS	/2012 0:00:00	1804.00	57.00	1677.0	15.0
28319		177	SS	2/2012 0:00:00	2106.00	54.00	20 .00	4.00
28320		177	SS	1/12/2012 0:00:00	1816.00	57.00	1687.00	5.00
28321	TOTALE					0	0	

È un problema di rappresentazione dei numeri da cui discende un problema con il formato dei dati.

Data type (II)


Formatta celle
Protezione celle
Numeri
Carattere
Effetti carattere
Allineamento
Tipografia asiatica
Bordo
Sfondo
Categoria
Numero
Percentuale
Valuta
Data
Orario
Scientifico
Frazione
Valore booleano
Formato
General
-1234
-1234.12
-1,234
-1,234.12
-1,234.12
(1,234)
(1,234.12)
Lingua
Inglese (USA)
Inglese (USA)
Inglese (Zimbabwe)
Interlingua
Irlandese
Islandese
Italiano (Italia)
Italiano (Svizzera)
Opzioni
Posizioni decimali
2
Valori negativi in rosso
Zeri iniziali
1
Separatore delle migliaia
Codice del formato
0.00
OK
Annulla
?
Ripristina

Data type (III)



Altri trucchi

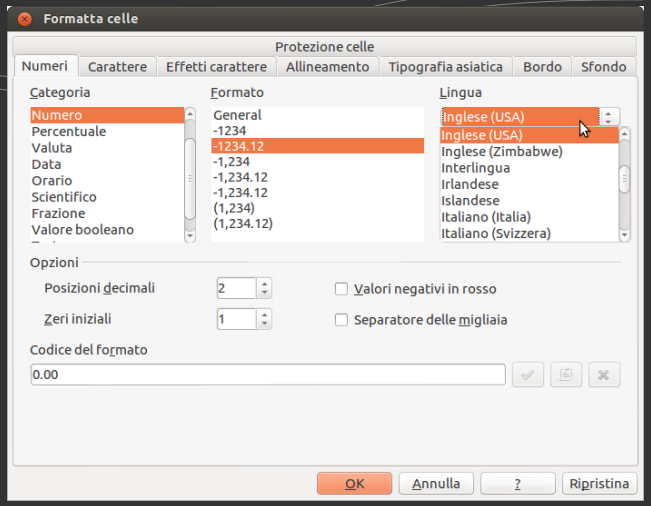
Modifica/Trova e sostituisci

Espressione regolare:
trova: ^.*$ → sostituisci: &

Applicare ai valori.

È possibile poi tornare alla lingua italiana (“.” → “,”)

(Oppure si può importare direttamente con l'impostazione in inglese)



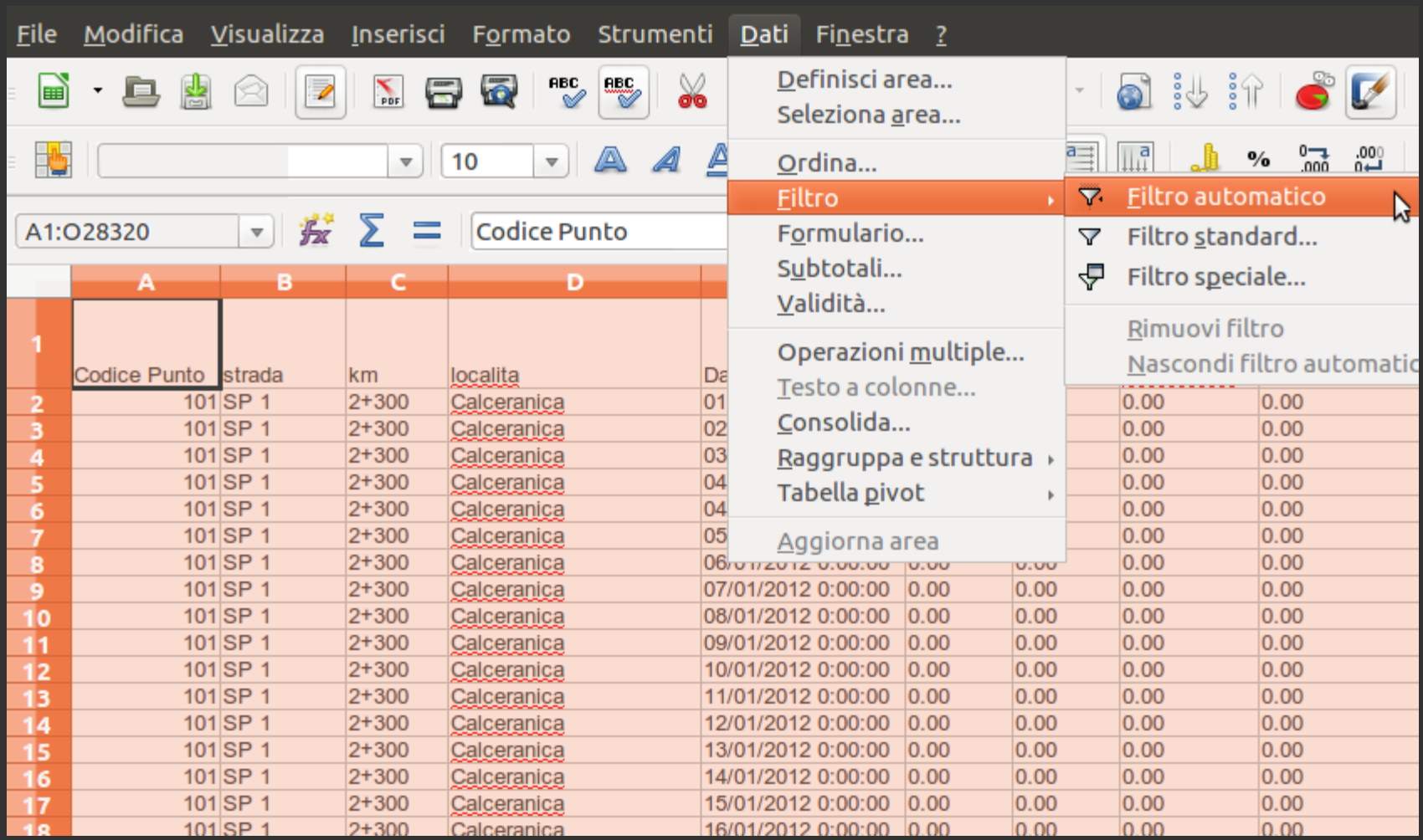
Espressioni regolari

http://www.regular-expressions.info/tutorial.html

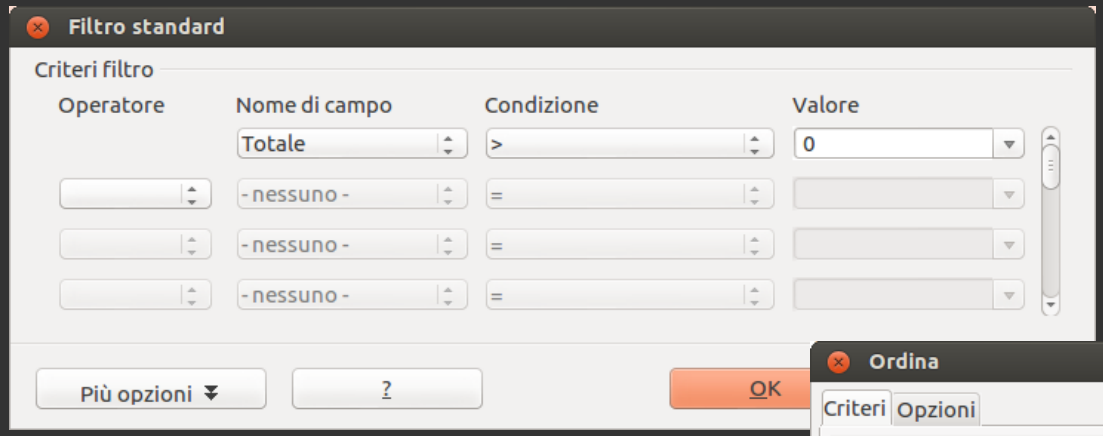
«Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems.»

Jamie Zawinski, alt.religion.emacs
(http://en.wikiquote.org/wiki/Jamie_Zawinski)

Filtraggio dei dati

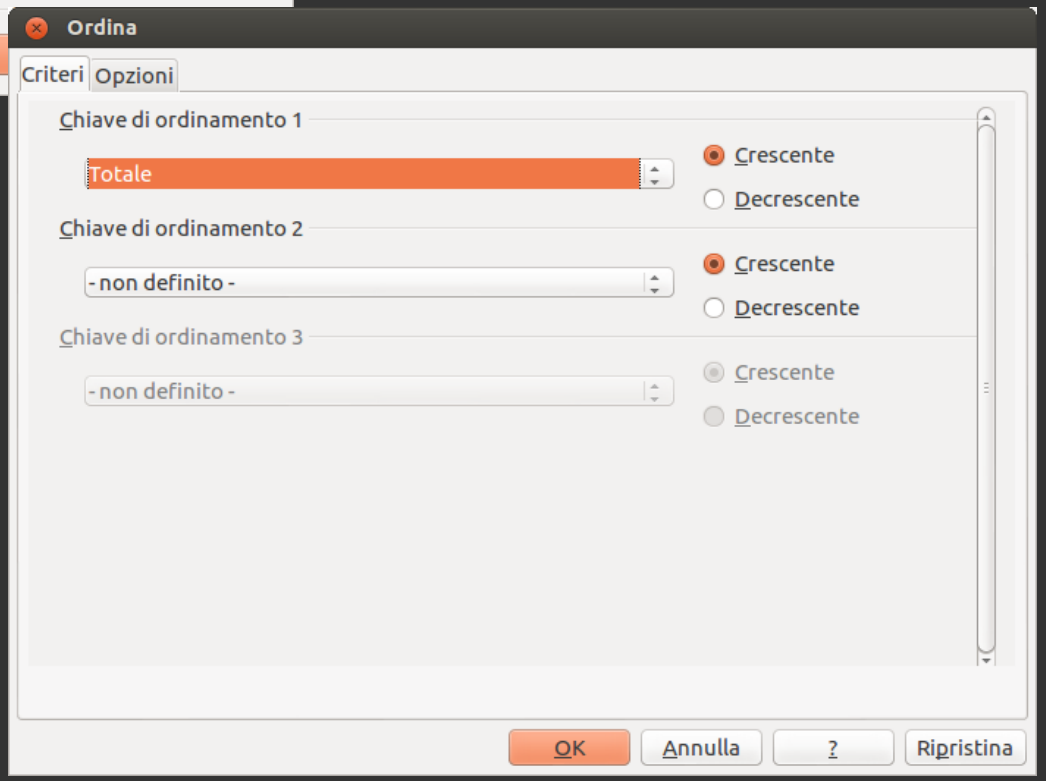


Filtro e ordinamento

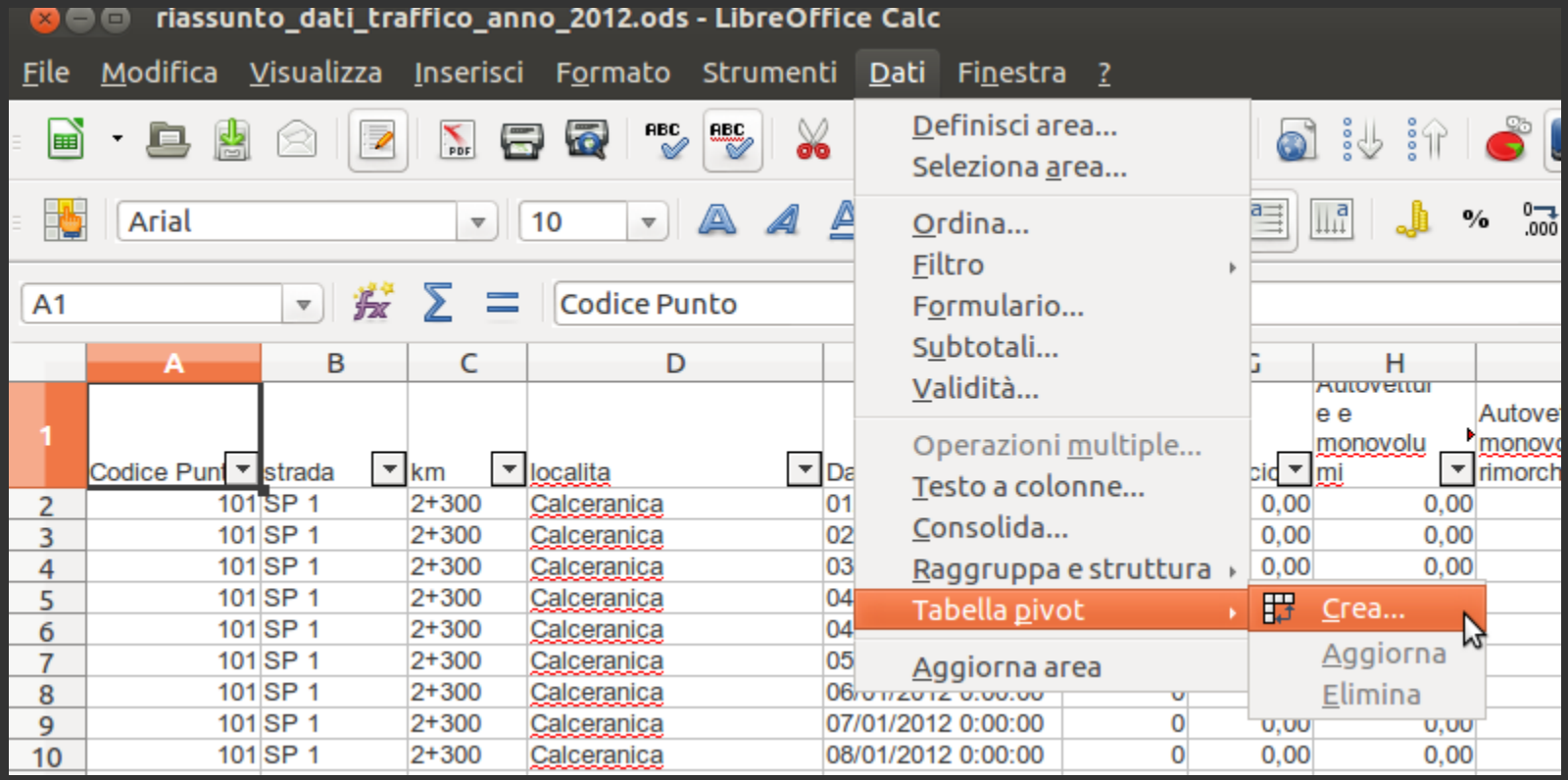


Filtri condizionali dei dati

Dati/Ordina ...



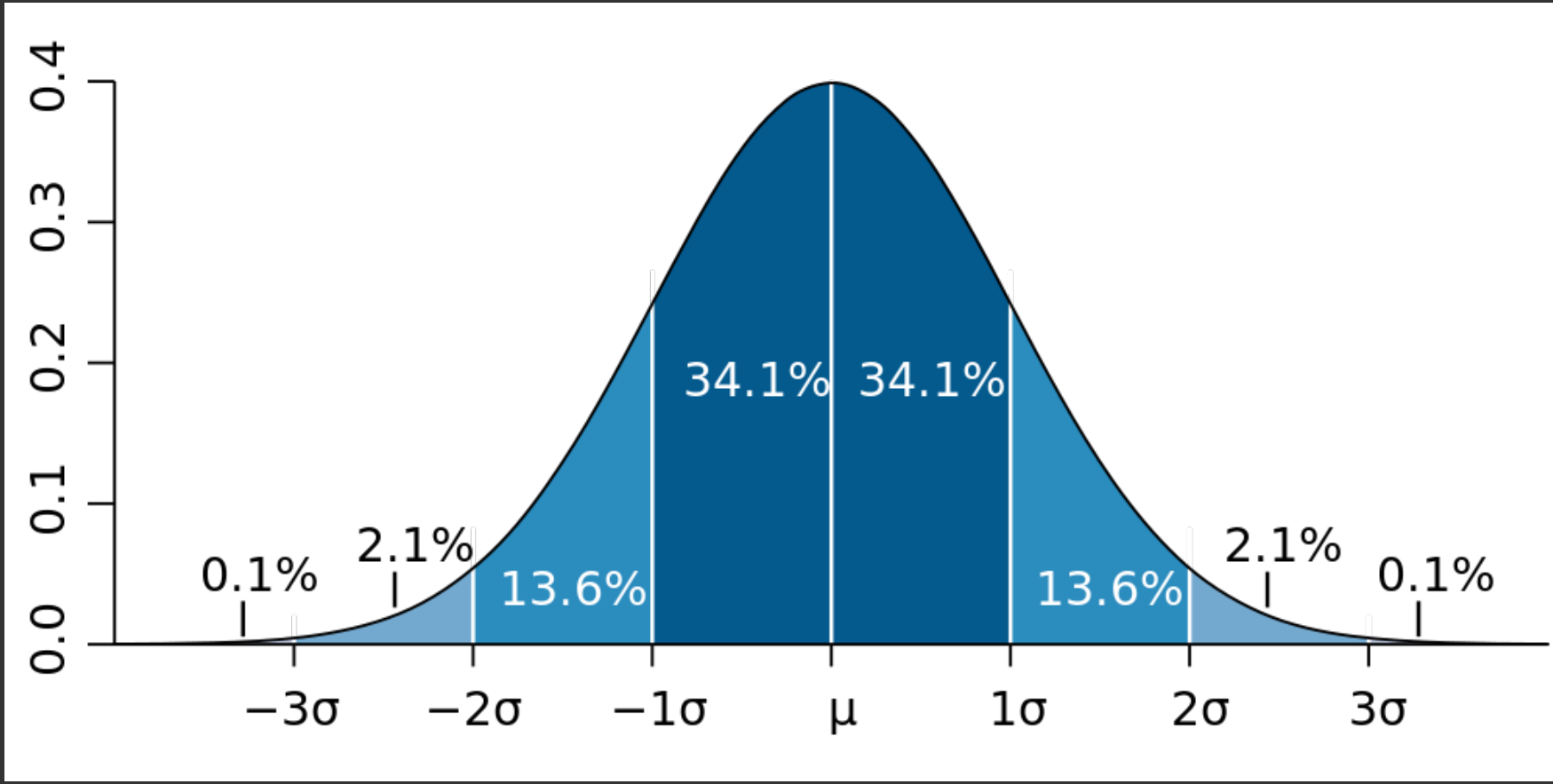
Acquisizione: pivot tables



Funzioni di base

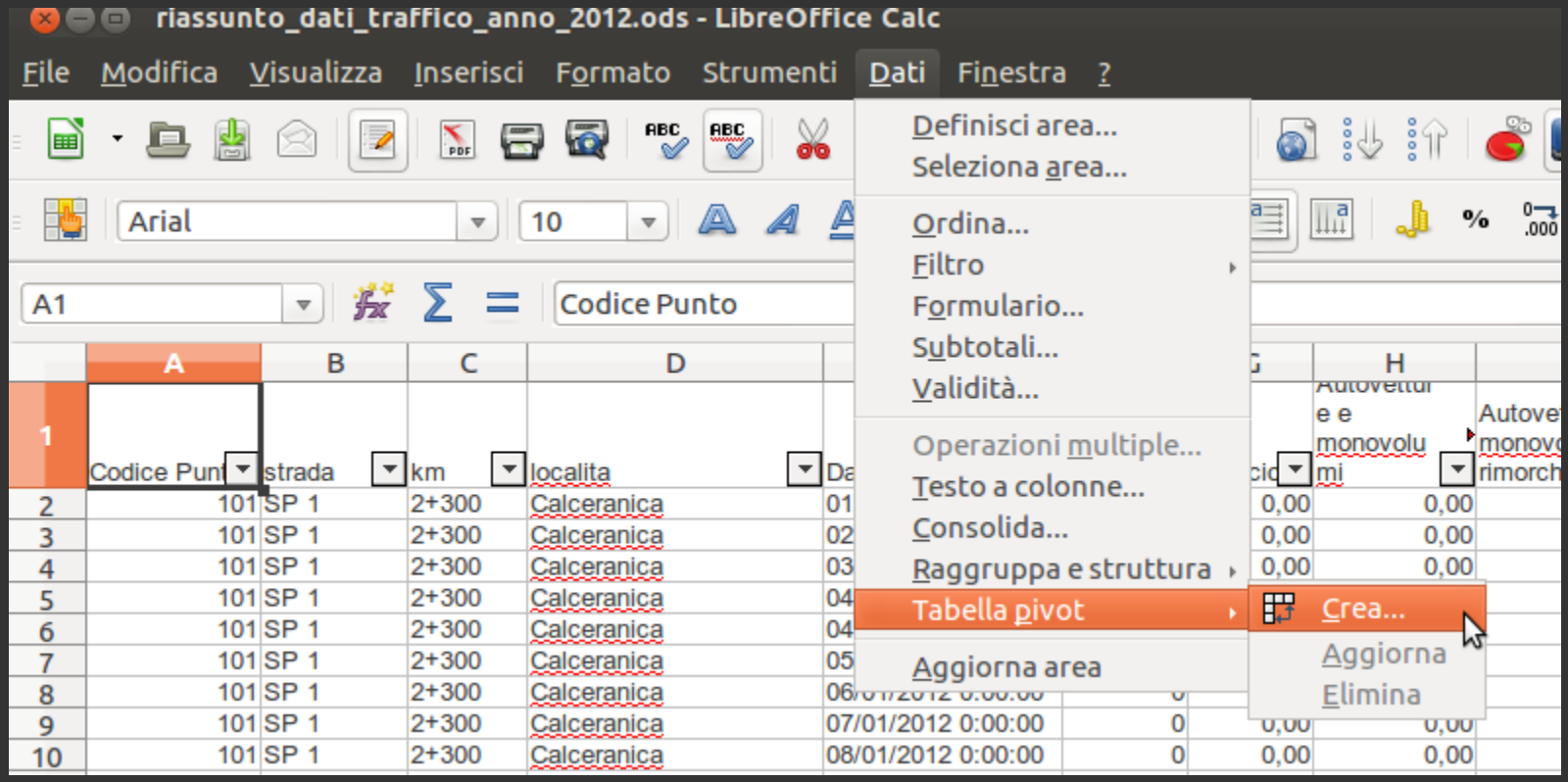
- Matematiche
  - SOMMA
  - MEDIA
  - CONTA.SE
- Logiche
  - SE
- Testo
  - CONCATENA
  - STRINGA.ESTRAI
- Statistiche
  - DEV.ST.POP

(intermezzo statistico)

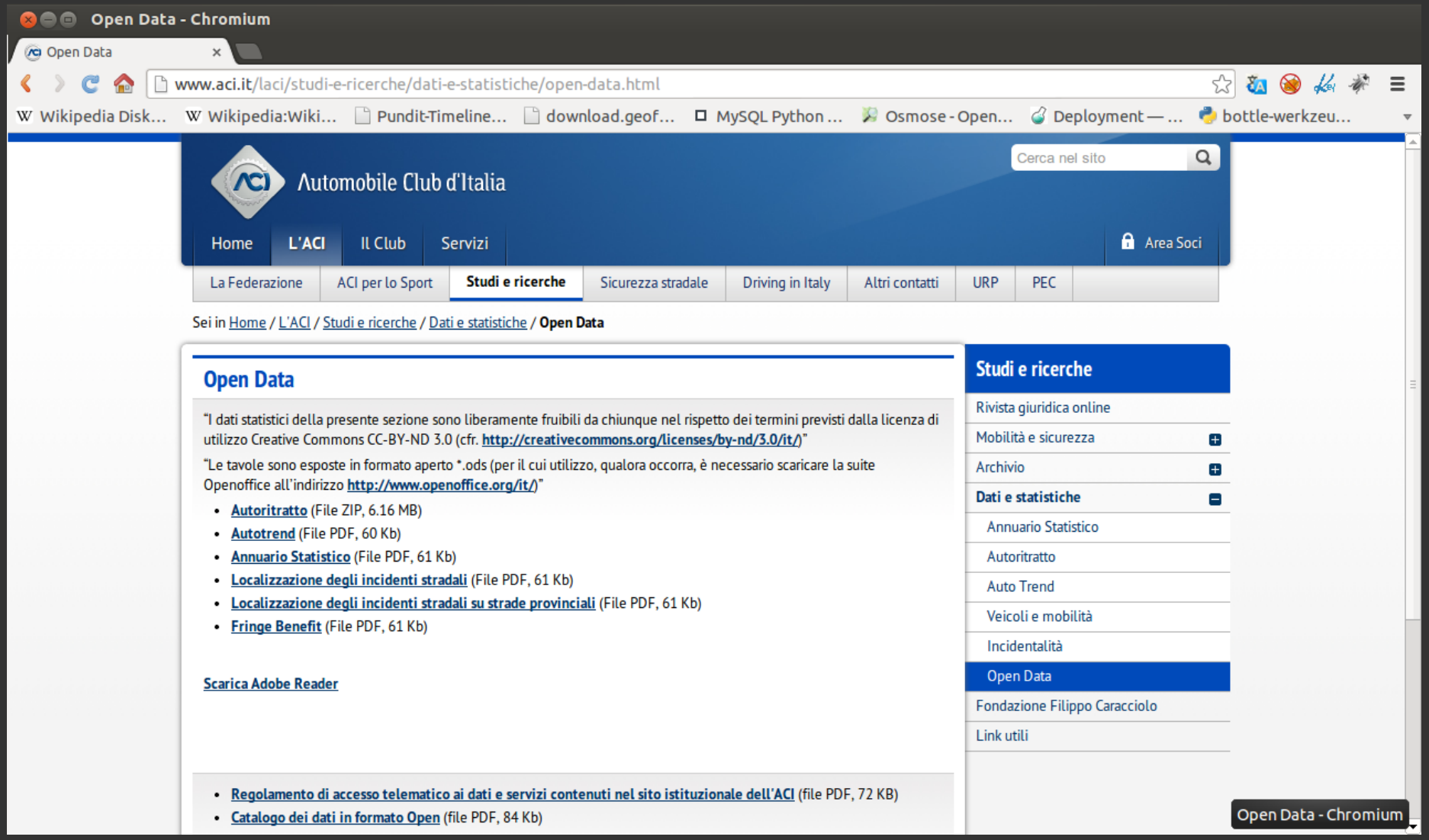


https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg

Tabelle pivot



(intermezzo sull'orgine dei dati)



(intermezzo sull'orgine dei dati)

Open Data - Chromium

Open Data

www.aci.it/laci/studi-e-ricerche/dati-e-statistiche/open-data.html

Wikipedia Disk... Wikipedia:Wiki... Pundit-Timeline... download.geof... MySQL Python ... Osmose - Open

Automobile Club d'Italia

Home | L'ACI | Il Club | Servizi

La Federazione | ACI per lo Sport | Studi e ricerche | Sicurezza stradale | Driving in Italy | Altri contatti | URP | PE

Sei in Home / L'ACI / Studi e ricerche / Dati e statistiche / **Open Data**

Open Data

"I dati statistici della presente sezione sono liberamente fruibili da chiunque nel rispetto dei termini previsti dalla licenza di utilizzo Creative Commons CC-BY-ND 3.0 (cfr. http://creativecommons.org/licenses/by-nd/3.0/it/)"

"Le tavole sono esposte in formato aperto *.ods (per il cui utilizzo, qualora occorra, è necessario scaricare la suite Openoffice all'indirizzo http://www.openoffice.org/it/)"

- Autoritratto (File ZIP, 6.16 MB)
- Autotrend (File PDF, 60 Kb)
- Annuario Statistico (File PDF, 61 Kb)
- Localizzazione degli incidenti stradali (File PDF, 61 Kb)
- Localizzazione degli incidenti stradali su strade provinciali (File PDF, 61 Kb)
- Fringe Benefit (File PDF, 61 Kb)

Scarica Adobe Reader

- Regolamento di accesso telematico ai dati e servizi contenuti nel sito istituzionale dell'ACI (file PDF, 72 KB)
- Catalogo dei dati in formato Open (file PDF, 84 Kb)

Studi e ricerche

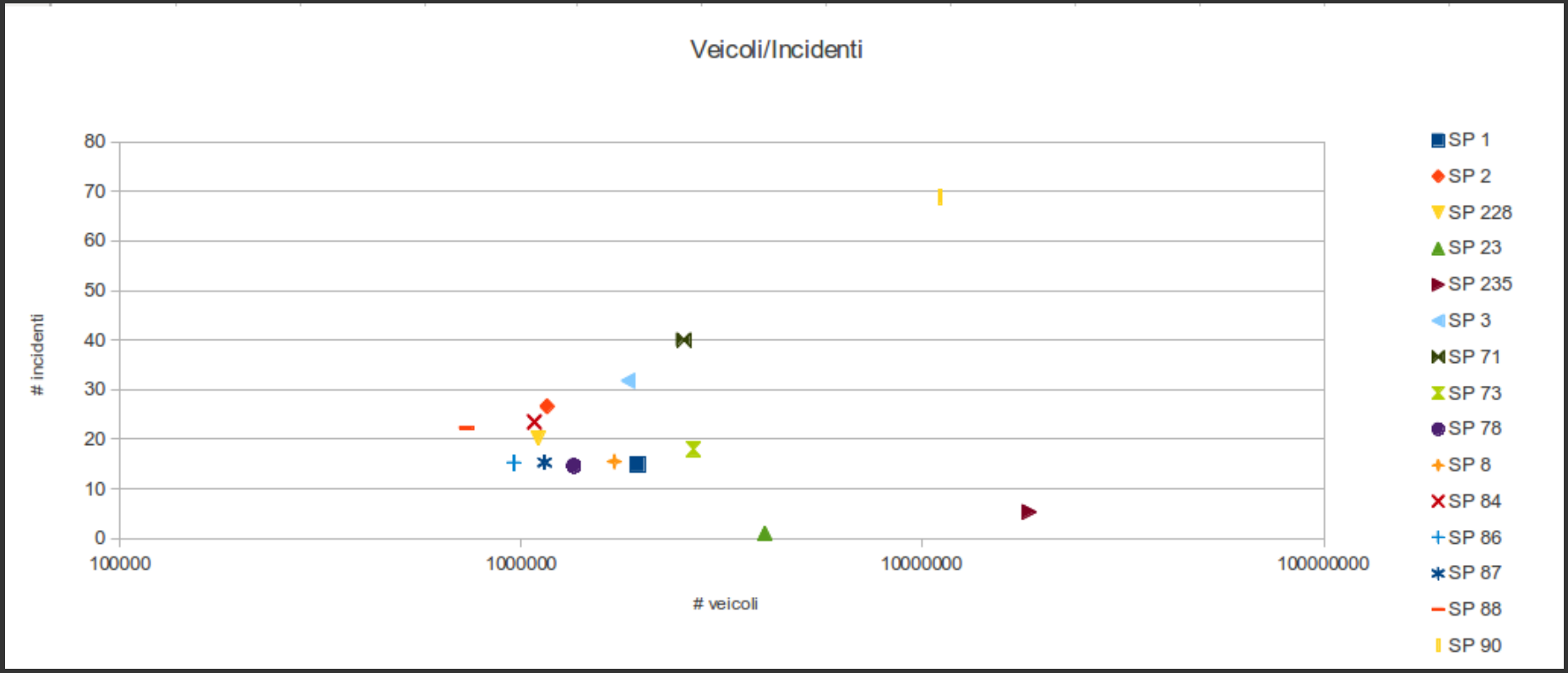
- Rivista giuridica online
- Mobilità e sicurezza
- Archivio
- **Dati e statistiche**
  - Annuario Statistico
  - Autoritratto
  - Auto Trend
  - Veicoli e mobilità
  - Incidentalità
  - Open Data
- Fondazione Filippo Caracciolo
- Link utili

Open Data - Chromium

#incidenti vs # veicoli

Nome strada	Veicoli	Incidenti
SP 1	1954918	14,866
SP 2	1162140	26,605
SP 228	1104131	20,217
SP 23	4051961	0,932
SP 235	18431427	5,3
SP 3	1854137	31,779
SP 71	2540191	40
SP 73	2690400	17,923
SP 78	1354838	14,592
SP 8	1709950	15,49
SP 84	1080852	23,457
SP 86	963420	15,207
SP 87	1147427	15,28
SP 88	731765	22,2
SP 90	11073415	68,868

un grafico


Veicoli/Incidenti
incidenti
80
70
60
50
40
30
20
10
0
100000
1000000
10000000
100000000
veicoli
SP 1
SP 2
SP 228
SP 23
SP 235
SP 3
SP 71
SP 73
SP 78
SP 8
SP 84
SP 86
SP 87
SP 88
SP 90

Data analysis: challenges

- Attenzione ai numeri piccoli
- Attenzione agli eventi rari
- Quali sono gli andamenti di lungo termine?
- Non lasciatevi trasportare dalle percentuali.
- Non lasciatevi trasportare dai numeri “ad effetto”

«*The lesson from this is if it sound ridiculous, it probably is, and it needs to be checked thoroughly, which is not the easiest thing to do when you are on deadline.*»

“Getting started with data journalism”, Claire Miller

Come salvare i propri dati

- Usare colori o strani font è inutile: *non fatelo!*
- È possibile esportare in CSV → nessun problema di compatibilità;
  - Si salva solo il foglio attivo
  - Non si salvano le formule o la formattazione!
- Utilizzando le funzionalità base (e salvando ne “vecchio” formato .xls, nel caso di Excel [97, 2000, XP, 2003], si riducono i problemi di compatibilità.
- Con formati aperti i problemi di compatibilità non si pongono! → I formati aperti sono *future proof*

Data cleaning: l'obiettivo

I dati devono essere spesso puliti per essere resi *omogenei*.

- Fase di preparazione dei dati
- Permette di creare visualizzazioni facilmente
- È un ottimo momento per iniziare a dare un'occhiata ai dati nel dettaglio

Raccolta di (alcuni) strumenti avanzati

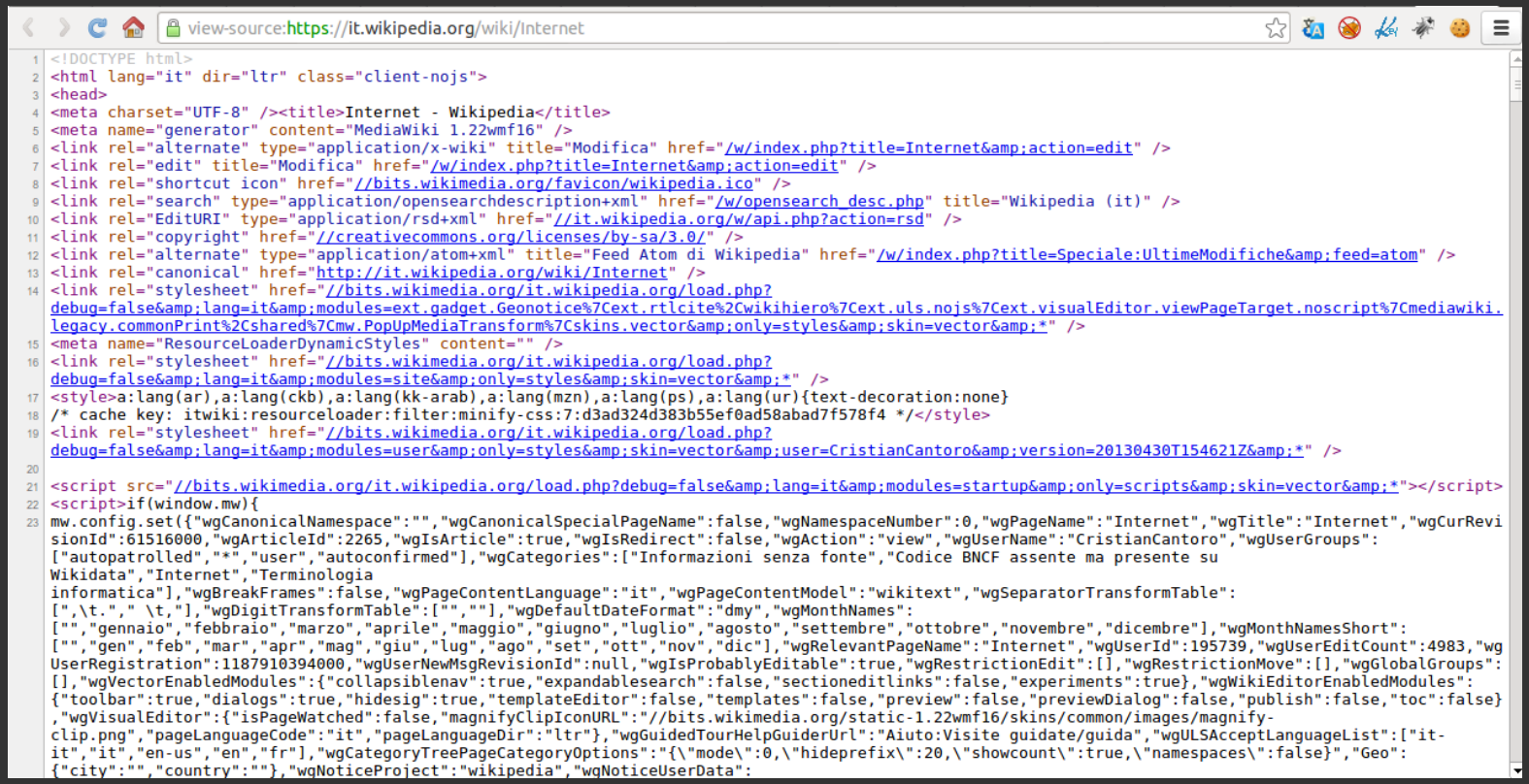
- Raw http://raw.densitydesign.org/
- Datawrapper http://datawrapper.de/
- Google Fusion Tables http://tables.googlelabs.com/
- Geojson.io http://geojson.io/

Scraping (I): in generale

- A volte basta un semplice copia-incolla
- Se la pagina è strutturata è relativamente semplice.
- Si può considerare l'ipotesi di pagare un programmatore per ottenere i dati (“outsourcing”).

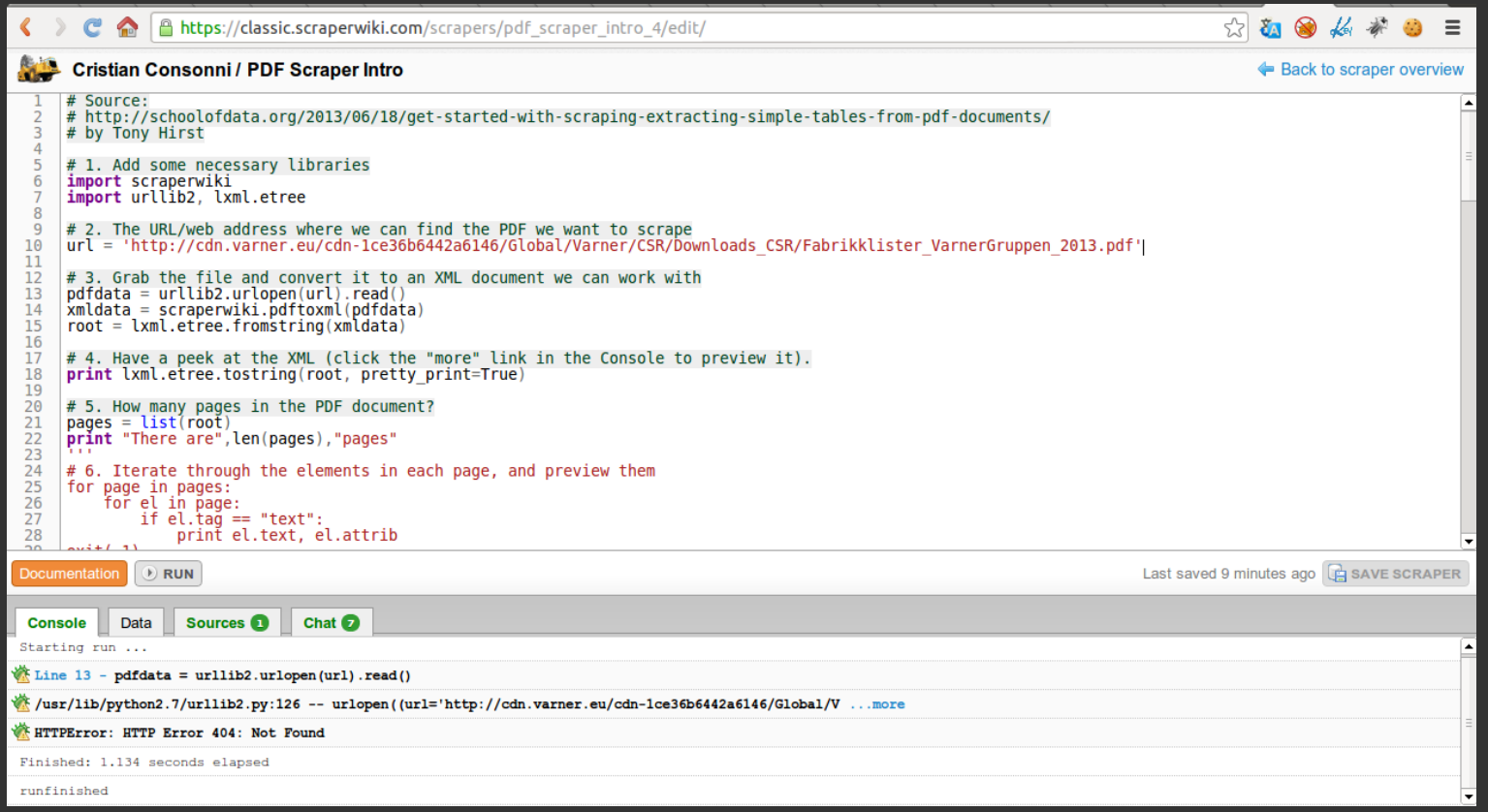
Scraping (I)

Sorgente HTML di una pagina:



Scraping (III): strumenti avanzati

ScraperWiki



Scraping (IV)

PDF:

«Scraping PDFs is a bit like cleaning drains with your teeth. It's slow, unpleasant, and you can't help but feel you're using the wrong tools for the job. […] Why is scraping PDFs so hard? Well, the PDF standard was designed to do a particular job: describe how a document looks, anywhere and forever.»

Tratto da:
http://blog.scraperwiki.com/2010/12/17/scraping-pdfs-now-26-less-unpleasant-with-scraperwiki/

Tutorial per chi vuole cimentarsi con un po' di codice:
http://schoolofdata.org/2013/06/18/get-started-with-scraping-extracting-simple-tables-from-pdf-documents/

Quali sono i rischi quando si lavora con i dati

✗ le teorie si adattano ai dati, non viceversa.

«Se le realtà non si adatta alla teoria, la realtà è sbagliata,»
(a volte erroneamente attribuita a Einstein)

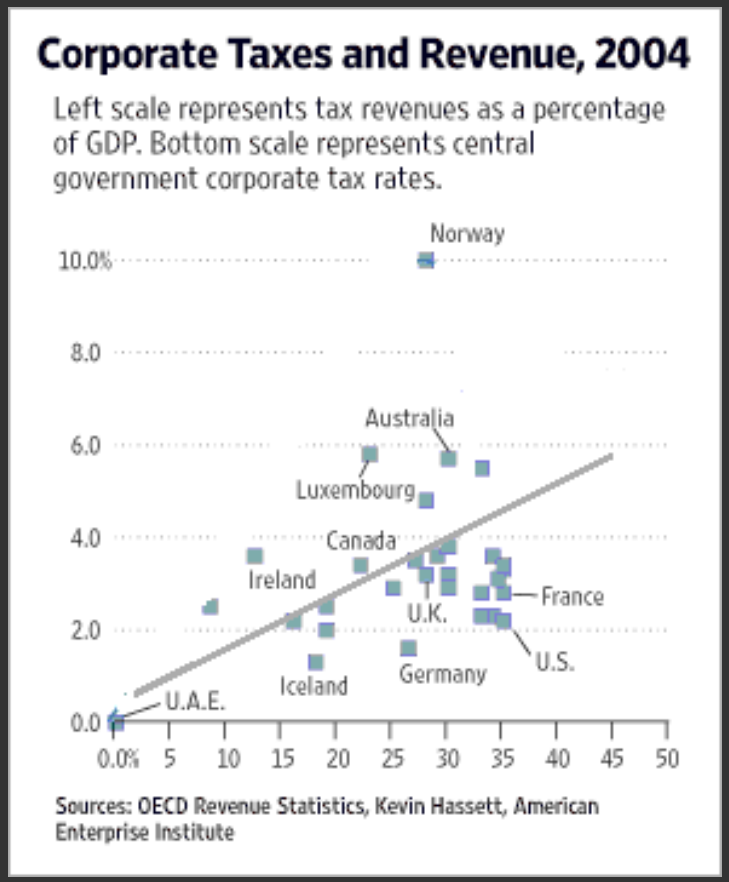
✗ correlazione non implica causalità.

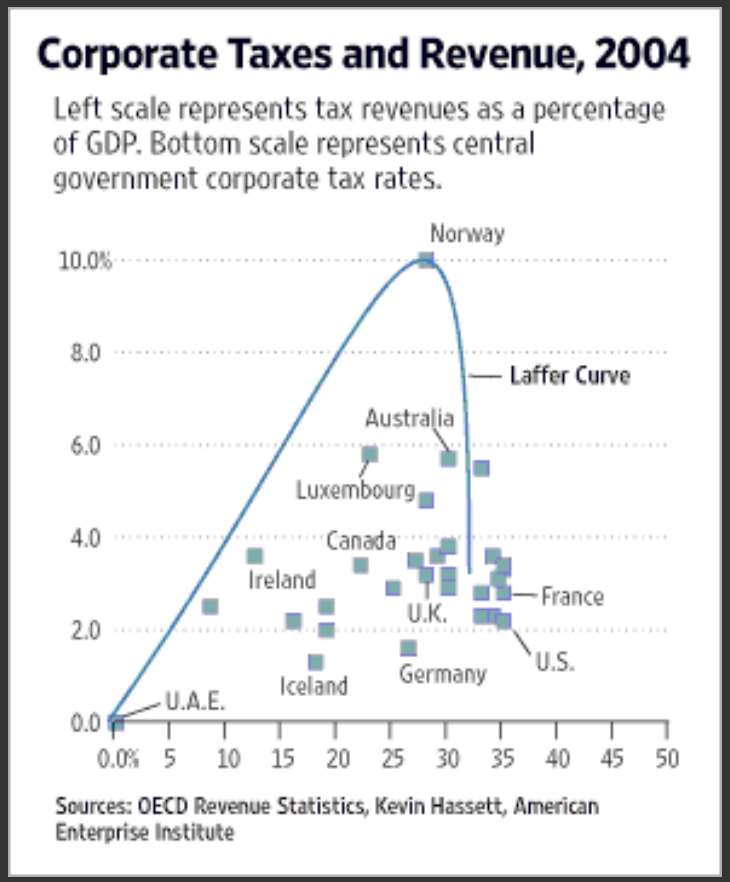
«Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.»
http://xkcd.com/552

✗ i modelli teorici sono sempre validi entro certi limiti.

«Finché le leggi della matematica si riferiscono alla realtà, non sono certe, e finché sono certe, non si riferiscono alla realtà,»
Albert Einstein, *Sidelights on Relativity*

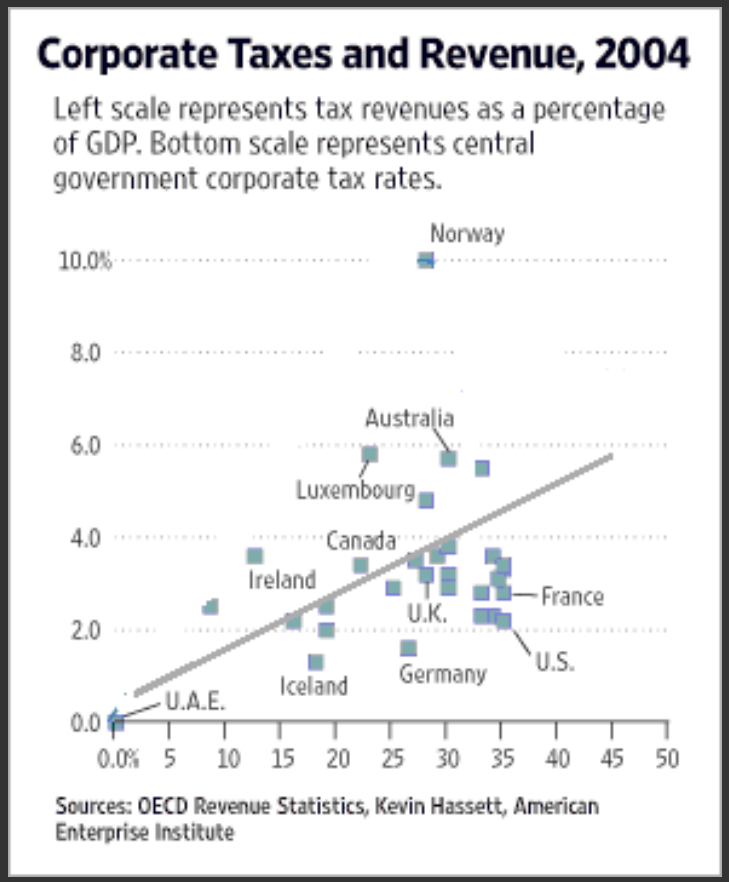
Rischi (1): adattare i dati alla teoria

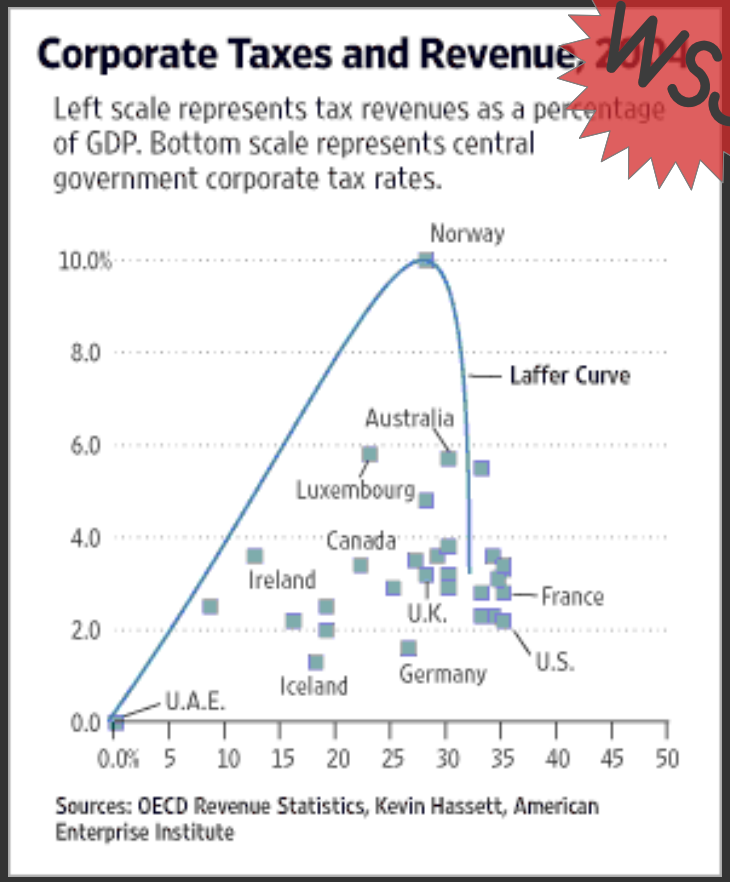




www.preposterousuniverse.com/blog/2007/07/13/the-best-curve-fitting-ever/

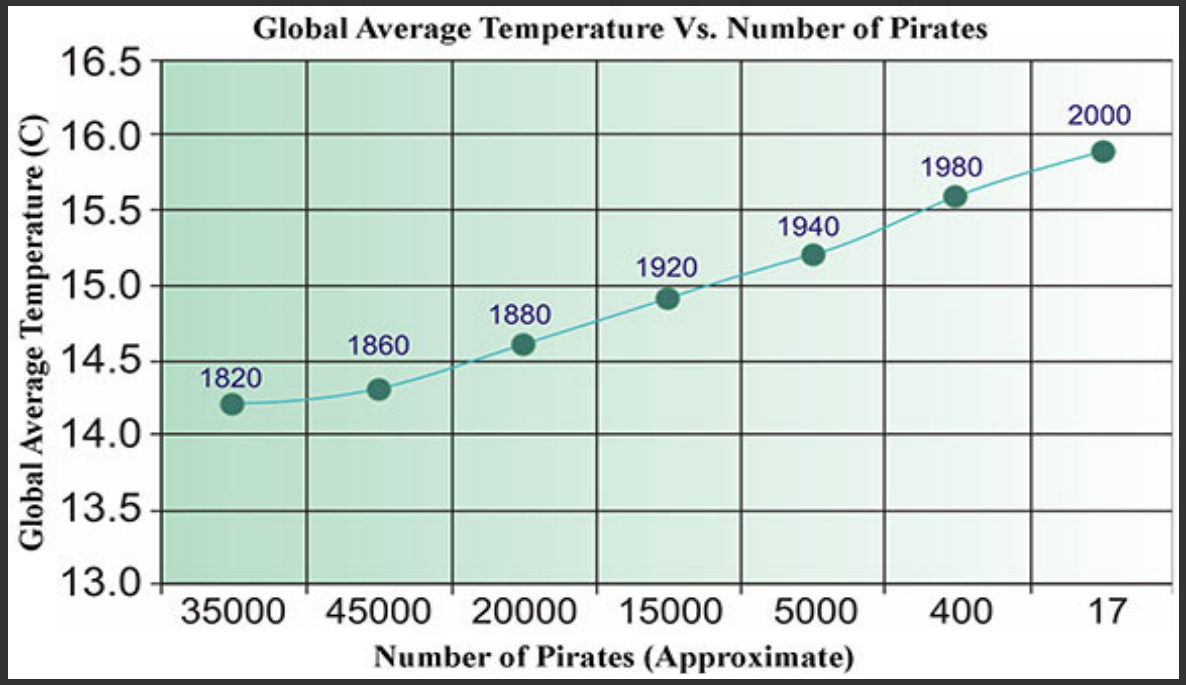
Rischi (1bis): adattare i dati alla teoria





www.preposterousuniverse.com/blog/2007/07/13/the-best-curve-fitting-ever/

Rischi (2): correlazione → causalità? No!



http://bressanini-lescienze.blogautore.espresso.repubblica.it/2013/02/15/mangia-cioccolato-e-vinci-il-premio-nobel/

FONDAZIONE
BRUNO KESSLER

Cristian Consonni

Mail: consonni@fbk.eu

CristianCantoro →

{ skype, twitter, wiki*, slideshare, ...}

SCHOOL OF DATA

FONDAZIONE BRUNO KESSLER

Find this presentation on slideshare:
http://www.slideshare.net/CristianCantoro

Credits

Questa presentazione è abbondantemente inspirata a quella di Marco Montanari:

- http://www.slideshare.net/sirmmo/rcs-27211305

Questa presentazione è rilasciata con licenza

CC-BY-SA

- http://creativecommons.org/licenses/by-sa/3.0/deed.it

Credits