Buone pratiche per lavorare con dati e metadati

software e strumenti per manipolare dati testuali

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Piero Grandesso, Chiara Festa, Giacomo D'Attorre

19 ottobre 2023
ARPAC – Settore Gestione e sviluppo della biblioteca digitale di Ateneo – AlmaDL

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Introduzione

scaletta

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

09:30–09:40	Introduzione e panoramica *Library Carpentry*
09:40–10:10	Gestione dei dati, buone pratiche
10:10–10:40	Espressioni regolari (*RegEx*)
10:40–10:55	*pausa* ☕
10:55–11:20	Espressioni regolari – pratica
11:20–12:00	OpenRefine
12:00–12:15	*pausa* ☕
12:15–13:30	OpenRefine – pratica

Disclaimer

- questo corso < *Library Carpentry*
  https://librarycarpentry.org/
- edizione "*crash course*"

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Regole del gioco

1. *panico?* 😱 → cartellino! 🆘
   se vediamo il post-it veniamo ad aiutarvi

2. pausa 🍪 → puntualità
   c'è davvero poco tempo

3. dialoghiamo, anche in modalità asincrona
   non scompariamo dopo il corso

Principi

- strumenti liberi (*free software*)
- documentazione accessibile (*open access*)
- condivisione di competenze (https://stackoverflow.com/)
- formati interoperabili
- ***F**indable, **A**ccessible, **I**nteroperable, **R**eusable*...

  suona già sentito?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

«Che c'entro io con l'informatica?»

L'**informatica** è la scienza che si occupa del **trattamento dell'informazione** mediante procedure automatizzate.
[wikipedia: Informatica]

Una **biblioteca** è un servizio finalizzato a soddisfare bisogni **informativi** [...] realizzato sulla base di una **raccolta organizzata** di supporti delle **informazioni**, fisici [...] o digitali.
[wikipedia: Biblioteca]

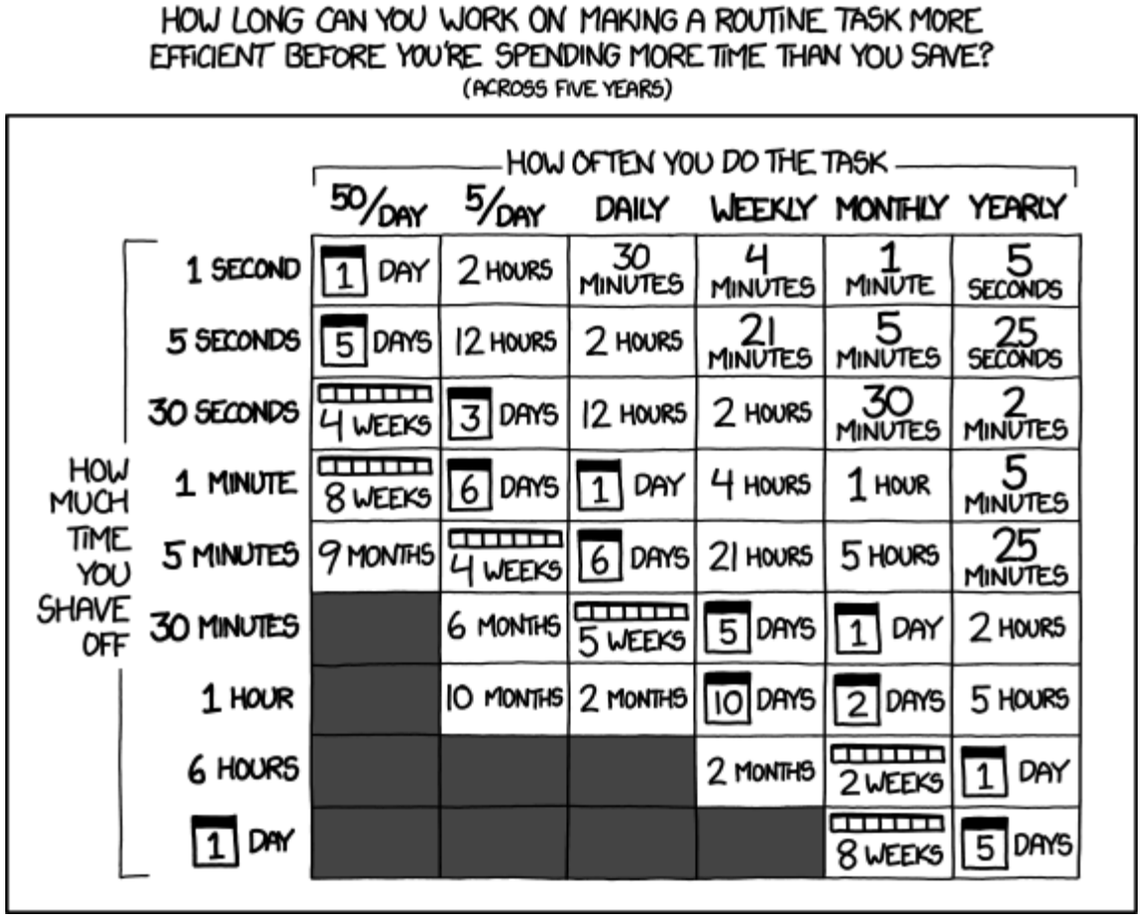
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Altri buoni motivi...

- automazione di processi ripetitivi
- riduzione del rischio di errori
- facilitare lo scambio di dati tra collegh[ei]
- collaborare con più efficacia con gli informatici

Is It Worth the Time?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



[https://xkcd.com/1205/ by Randall Munroe, CreativeCommons BY-NC 2.5]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Uno sguardo nel mondo di *Library Carpentry*

La *shell*

- la *shell Unix*, è un potente strumento per manipolare i file
- presente su computer Linux e Mac ma anche installabile su Windows[*]
- fondamentale per grandi quantità di file, file molto pesanti, file di testo semplice
- comandi ripetitivi possono essere raggruppati in *script* ripetibili al bisogno, possono includere variabili, condizioni e altre cose per renderli più flessibili

[*]: ad esempio https://gitforwindows.org/

Git

Git è un famoso software per il controllo della versione dei file, serve per:

- tenere traccia delle modifiche (e backup) dei file di un progetto
- permettere di annullare modifiche (*rollback*) o creare più rami di un progetto (*branches*)
- collaborare tra più utenti nella modifica e scrittura
- rende trasparente lo sviluppo del software che potreste utilizzare

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

SQL

Structured Query Language è un linguaggio per interrogare e manipolare database relazionali.

Un database relazionale è un insieme di tabelle (con colonne di dati omogenei e righe contenenti i vari record) che contengono dati messi in relazione tra loro ed è la base dei dati standard per quasi tutti gli applicativi web

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Gestione dei dati, alcune buone pratiche

Prendiamo una tabella...

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Titolo	Autore
Recueil d'observations faites en plusieurs voyages par ordre de sa majesté...	Cassini, Giovanni Domenico; Picard, Jean; Richer, Jean
De gnomone meridiano Bononiensi ad divi Petronii deque observationibus astronomicis eo instrumento...	Manfredi, Eustachio
La meridiana del tempio di San Petronio rinnovata l'anno 1776	Cassini, Giovanni Domenico; Guglielmini, Domenico
Ephemerides nouisssimae motuum coelestium marchionis Cornelii Maluasiae ...	Malvasia, Cornelio; Cassini, Giovanni Domenico; van Lansberge, Philippe; Stringa, Francesco

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

...non tutti i formati sono uguali

(qui e altrove siamo usciti dalle slide per guardare direttamente gli esempi...)

Plain text – vantaggi

- minore complessità
- gestibili con una molteplicità di strumenti, lo saranno anche nel futuro
- minori distrazioni (puro contenuto)

Editor di testo

Programmi pensati per visualizzare e modificare file di testo semplice:

- Notepad++ (windows)
- Visual Studio Code (windows, mac, linux)
- TexEdit (mac – ma occhio alla modalità)
- **non** Word **né** il blocco note!

Editor di testo – feature

- contatore di righe e caratteri
- funzione di Trova/Sostituisci avanzata (regex)
- evidenziazione della sintassi (se scrivete XML o Markdown il testo si colora per facilitare la lettura)
- scorciatoie, estensioni, correttori automatici, suggeritori di parola etc

Formati a testo semplice

Puri: TXT, CSV (_comma-separated values_), TSV (_tab-separated values_)

Strutturati (o annotati): XML, HTML, Markdown, JSON, TeX

Formati a testo semplice

Puri: TXT, CSV (*comma-separated values*), TSV (*tab-separated values*)

Strutturati (o annotati): XML, HTML, Markdown, JSON, TeX

I caratteri

Esistono varie codifiche e una buona dose di complessità, anche nel testo semplice...

- ASCII

```
 !"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
```

I caratteri – set estesi

Si crea una babele di codifiche differenti, quasi mai compatibili tra loro...[*]

- vecchi: ISO/IEC 8859 e derivati (Windows-1252 per l'alfabeto latino)
- esotici: Shift JIS, GB 2312 ...
- e nuovi: **Unicode (UTF-8 e UTF-16)**

[*]: se vi interessa l'argomento https://www.youtube.com/watch?v=MijmeoH9LT4

Interruzione riga

Persino su questo non esiste uno standard:

- Windows: 2 caratteri, CR+LF, il *carriage return* e il *line feed* [`\r\n`]
- Linux e MacOS: 1 carattere, LF [`\n`]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Machine readable

Facilitiamo la vita alle macchine (ma pure a noi!) adottando gli standard esistenti: `en-US` o `CL` è molto più gestibile e neutro rispetto alla lingua usata che scrivere “inglese americano” o “Cile”.

Occhio agli standard esistenti, in particolare nelle sigle di lingue (ISO 639) e paesi (ISO 3166) e nelle date (ISO 8601).

- per esempio su Historica abbiamo adottato per le lingue l’ISO 639-2

Gli standard – le date

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

[https://xkcd.com/1179/ by Randall Munroe, CreativeCommons BY-NC 2.5]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Espressioni regolari

Definizione

Le espressioni regolari (o *regex*) sono sequenze di caratteri che consentono di definire una serie di stringhe (che combaciano con la sequenza: *match*). Definiscono un percorso di ricerca (*pattern*) all'interno di un testo, con molta più precisione rispetto alle *wildcard* spesso usate nei motori di ricerca (`* ? %`).

wildcard

`student*` può includere "studenti" e "studentesse", ma anche "studentato", "studentelli"...

regex

`student(i|esse)` è molto più preciso, non lascia spazio ad alternative, così come `bibliotecari[ao]` offre due sole opzioni.

Usi

Le *regex* per essere usate devono poter essere supportate dai nostri strumenti, per esempio:

- qualsiasi editor di testo semplice (degno del nome)
- LibreOffice / Open Office
- Open Refine
- Total Commander e AntRenamer
- una pletora di strumenti a riga di comando
- ma in realtà quasi ogni software avanzato per la gestione di dati
- ah no, niente Excel! 🤦

Principali regole

elemento seq.	match
`[abc]`	Un carattere tra i seguenti: `a`, `b`, `c`
`[^abc]`	Un carattere eccetto: `a`, `b`, `c`
`[a-z]`	Un carattere nell'intervallo: `a-z`
`[^a-z]`	Un carattere fuori dall'intervallo: `a-z`
`[a-zA-Z]`	Un carattere negli intervalli: `a-z`, `A-Z`
`.`	Un qualsiasi carattere
`\s`	Un carattere di spaziatura (orizz+vert)
`\S`	Un carattere eccetto quelli di spaziatura

Principali regole – 2

el. seq.	match
`\d`	Una cifra = `[0-9]`
`\D`	Un carattere eccetto le cifre
`\w`	*word-character* = `[a-zA-Z0-9]`
`\W`	Un *non-word-character*
`(...)`	*Capture group*
`(ab|cd)`	O la seq. “`ab`” oppure “`cd`”

Principali regole – 3

el. seq.	match
`a?`	Zero o un “a”
`a*`	Zero o più “a”
`a+`	Una o più “a”
`a{3}`	Esattamente 3 “a” = `aaa`
`a{3,}`	3 o più “a”
`a{3,6}`	Non meno di 3 e non più di 6 “a”
`^`	Inizio della stringa
`$`	Fine della stringa

regex – esempio

Cosa prende questa *regex*? `^199[5-7]$`

```
1993
2995
1995
1999
19 97
```

```
millenovecentonovantacinque
199
19957
1996
 1997
```

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Limiti

Non esiste uno standard unico per le *regex*, la cui sintassi e le cui possibilità variano quindi a seconda degli strumenti e del contesto.

I testi possono presentare codifiche differenti e caratteri di vario tipo, che non sempre le nostre *regex* sapranno trattare.

Soprattutto: attenzione a non *matchare* più del voluto!

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Esercizi!

materiali → **https://tinyurl.com/HistoricaOR-23**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

OpenRefine

Flusso ideale

È uno strumento per pulire e ordinare dati sporchi, perché raccolti a mano, non standardizzati e magari provenienti da varie fonti.

Consente di importarli → pulirli e uniformarli → esportarli affinché possano essere usati.

Ogni dataset inserito in OpenRefine è un progetto le cui modifiche vengono tracciate.

Installazione

Aggiorna con OpenRefine 3.7 e info su dove trovare la cartella dei progetti

1. https://openrefine.org/download.html → Windows kit

   serve Java, se non l'avete e non potete installarlo provate con

   Windows kit with embedded Java
2. scompattare lo zip in una cartella a vostra scelta
3. doppio click su `openrefine.exe`

   o se non funziona su `refine.bat`
4. si apre una finestra (nera) che potete ignorare,

   ma non chiudere!

   si aprirà il browser su http://127.0.0.1:3333/

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Scaletta

- Importare un progetto
- Interfaccia
- Facet/filtri
- Trasformazioni
- Crea colonna da progetto
- Clustering
- Reconciling
- Esportare un progetto

Importare un progetto

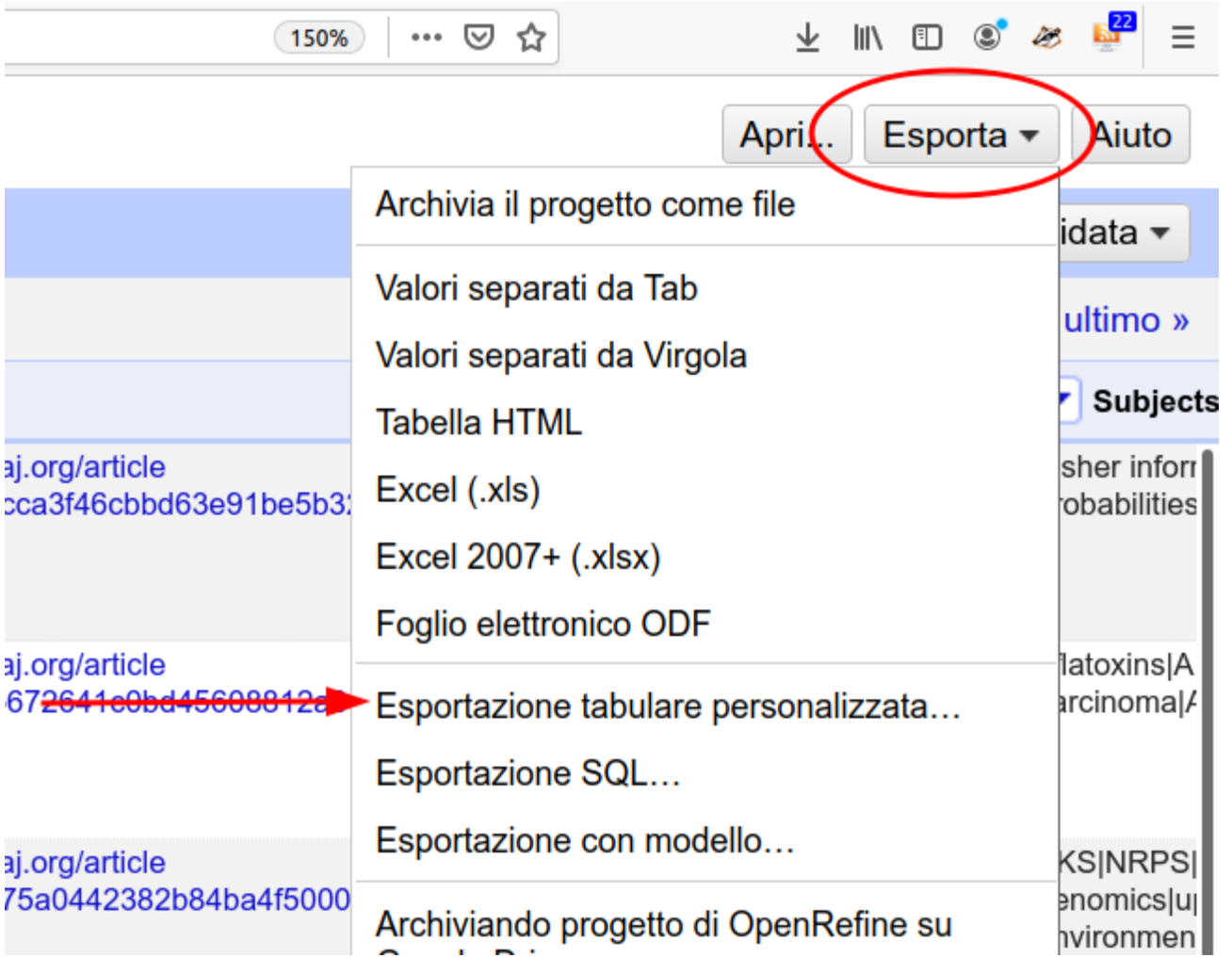
1. In “Crea progetto” varie opzioni:
   - carica da file (riconosce in automatico, potete aggiustare impostazioni, codifica etc)
   - appunti (copia/incolla)
   - link a Google Spreadsheet
   - …
2. Aggiustare le impostazioni, scegliere un nome e cliccare sul bottone “Crea progetto” (alto a dx)


OpenRefine doaj article sample
Permalink
Open...
Export
Help
Operazioni su colonna
Facet / Filter
Undo / Redo 12 / 15
1001 rows
Extensions: Wikidata
Refresh
Reset All
Remove All
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 10 next › last »
Language
change
4 choices Sort by: name count
Cluster
EN 871
English 107
ES 7
FR 1
(blank) 15
Facet by choice counts
Cronologia
Filtri e faccette
URL
Date
Language
Subjects
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom Numeric Facet...
Customized facets
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
e17127853
https://doaj.org/article/b75e8d5cca3f46cbb
nformation|quasi-ities|complement
agriculture5041172
https://doaj.org/article/0edc5af6672641c0b
s|AFM1|AFB1|m na|Agriculture (G
jms161226101
https://doaj.org/article/d9fe469f75a044238
PS|metagenomic s|upwelling|coas nent|Chemistry|C
norganics3040534
https://doaj.org/article/95606ed39deb4f43b96f7e6308ad15d3
01/11/2015
EN
lanthanide|cerium|scorpic non-innocent|Inorganic cl

Esportare un progetto

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Varie opzioni di esportazione, qui export flessibile in TSV oppure fogli Excel.


150%
Apri...
Esporta
Aiuto
Archivia il progetto come file
Valori separati da Tab
Valori separati da Virgola
Tabella HTML
Excel (.xls)
Excel 2007+ (.xlsx)
Foglio elettronico ODF
Esportazione tabulare personalizzata…
Esportazione SQL…
Esportazione con modello…
Archiviando progetto di OpenRefine su
idata
ultimo »
Subjects

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Fine

Grazie

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Piero Grandesso
<piero.grandesso2@unibo.it>

Chiara Festa
<chiara.festa4@unibo.it>

Giacomo D'Attorre
<giacomo.dattorre@unibo.it>

ARPAC – Settore Gestione e sviluppo della biblioteca digitale di Ateneo – AlmaDL

www.unibo.it