di:

Federica Belfiore

mat. 502041

Ludovica Pannitto

mat. 491094

INDICE

1. Introduzione

Il team del progetto è costituito dalle studentesse:

- Ludovica Pannitto (mat. 491094)
- Federica Belfiore (mat. 502041)

La presente relazione nasce con l'intento di presentare alcune considerazioni sul progetto "Uno studio del genere – Report sulle donne nell'Università italiana".

1.1 Obiettivo

Il progetto è stato pensato per realizzare una ricerca sulla distribuzione dell'organico nei vari atenei italiani in modo da analizzare la percentuale di docenti e ricercatori universitari in relazione a diverse variabili quali il sesso, l'area geografica e il settore di appartenenza. L'obiettivo di quest'analisi è quello di mettere in luce le disparità di genere ancora presenti nel mondo universitario e di promuovere politiche che favoriscano l'imparzialità di genere in tutti i campi di studio.

2. Stato dell'arte

Frutto di una ricerca sulla distribuzione dei docenti in tutti gli atenei italiani, che si è avvalsa di metodologie d'indagine quantitative e qualitative, l'articolo "Riequilibrare le opportunità nelle università e negli enti di ricerca" offre un'analisi dei problemi riscontrati negli ultimi anni nel mercato del lavoro all'interno dell'ambito universitario. Gli squilibri di opportunità, causati da stereotipi e discriminazioni di genere, riguardano prevalentemente il mondo femminile e sono il risultato di un complesso di rigidità sia sul versante della domanda sia da quello dell'offerta di lavoro. L'indagine qui presentata, oltre a fornire grafici rappresentativi dell'evoluzione nel tempo delle percentuali di docenti e della crescita dell'insegnamento universitario, offre i dati europei e italiani sui docenti, relativi al sesso, al grado di formazione, alla condizione occupazionale, alla formazione e all'area disciplinare. Ne emerge un percorso tortuoso che sfocia in un'accentuata diversificazione della condizione delle donne.

Quello che si profila, nonostante il numero di donne laureate, diplomate o dottoresse di ricerca sia maggiore rispetto a quello dei colleghi uomini, è che l'Italia è al 74esimo posto su 135 paesi, per la differenza di genere nel mondo del lavoro, e al 90esimo posto per "partecipazione economica ed opportunità": un dato preoccupante dipendente sia dalla mancanza totale di consapevolezza e informazione degli stessi universitari, sia da una più generale e complessa questione del lavoro femminile in Italia.

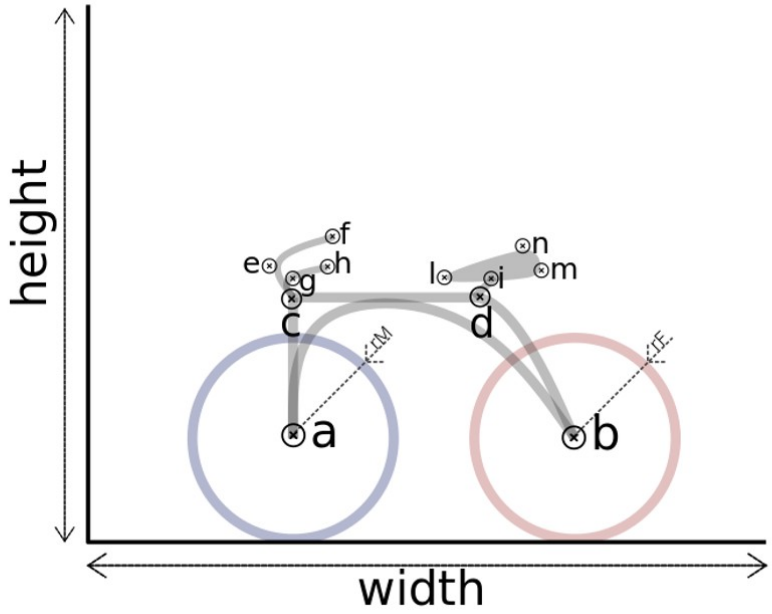
3. Illustrazione degli aspetti salienti del progetto

Esistono pochi siti web che dispongono di grafici, mappe e articoli scientifici in cui è resa disponibile un'analisi sull'argomento. Un sito web che vagamente riprende la tematica, dedicato più che altro all'ambito della ricerca, è http://chartsbin.com/view/1123. Uno degli aspetti più salienti di "Uno studio del genere" è sicuramente la presenza di dati statistici, relativi alla distribuzione dei docenti in Italia, aggiornati all'anno 2015: gli ultimi studi che hanno focalizzato l'attenzione sullo stesso campo risalgono al 2012. Viene inoltre utilizzato un database ampio e particolareggiato che permette l'implementazione di una ricerca il più possibile personalizzata, che raccoglie e analizza una considerevole quantità di dati e che concede la comparazione (con i dovuti limiti intrinsechi dipendenti dalla differenza tra le fonti di dati e dalla diversa struttura del database) tra quelli forniti dal CNR e quelli provenienti dall'università, pianificando così eventuali futuri requisiti di espansione. L'interfaccia che abbiamo scelto si rivolge ad un utente generico, tuttavia, nel caso in cui si vogliono sfruttare le potenzialità del database e si vogliano ricavare altri tipi di informazione, si può sempre scegliere di modificare adeguatamente la struttura del sito in funzione di uno specifico campione di utenti.

3.1 La visualizzazione

Trattandosi già di una delle tematiche più stereotipate, abbiamo voluto evitare di rappresentare il grafico concernente le percentuali di donne e uomini in modo banale e abituale, e abbiamo optato per una visualizzazione in cui l'argomento in questione non fosse immediatamente intuibile.

Così la scelta è ricaduta sul disegno di una bicicletta: il rapporto tra le aree delle ruote rispetta la proporzione tra il numero di docenti uomini e il numero di docenti donne. La costruzione è stata realizzata tramite la libreria d3 che ha permesso, a partire da pochi punti sul canovaccio svg, il riposizionamento degli stessi e dunque la rimodulazione dell'immagine ogni qual volta si esegua una ricerca.



Ogni punto è stato espresso tramite le sue coordinate sul piano

`R= heigth x width .`

(Nota: le ordinate risultano invertite: il valore 0 è il più alto del riquadro)

I punti (a) e (b) sono stati fissati in modo che le ruote della bicicletta toccassero sempre il fondo, secondo la legge:

```
{"y": height-5-rM ,"x": (width/2)+/-(rM + (4/10)*rM)}
```

A questo punto si è fissato:

```
(c)[y]=(a)[y]-( rM + (4/10)*rM)
```

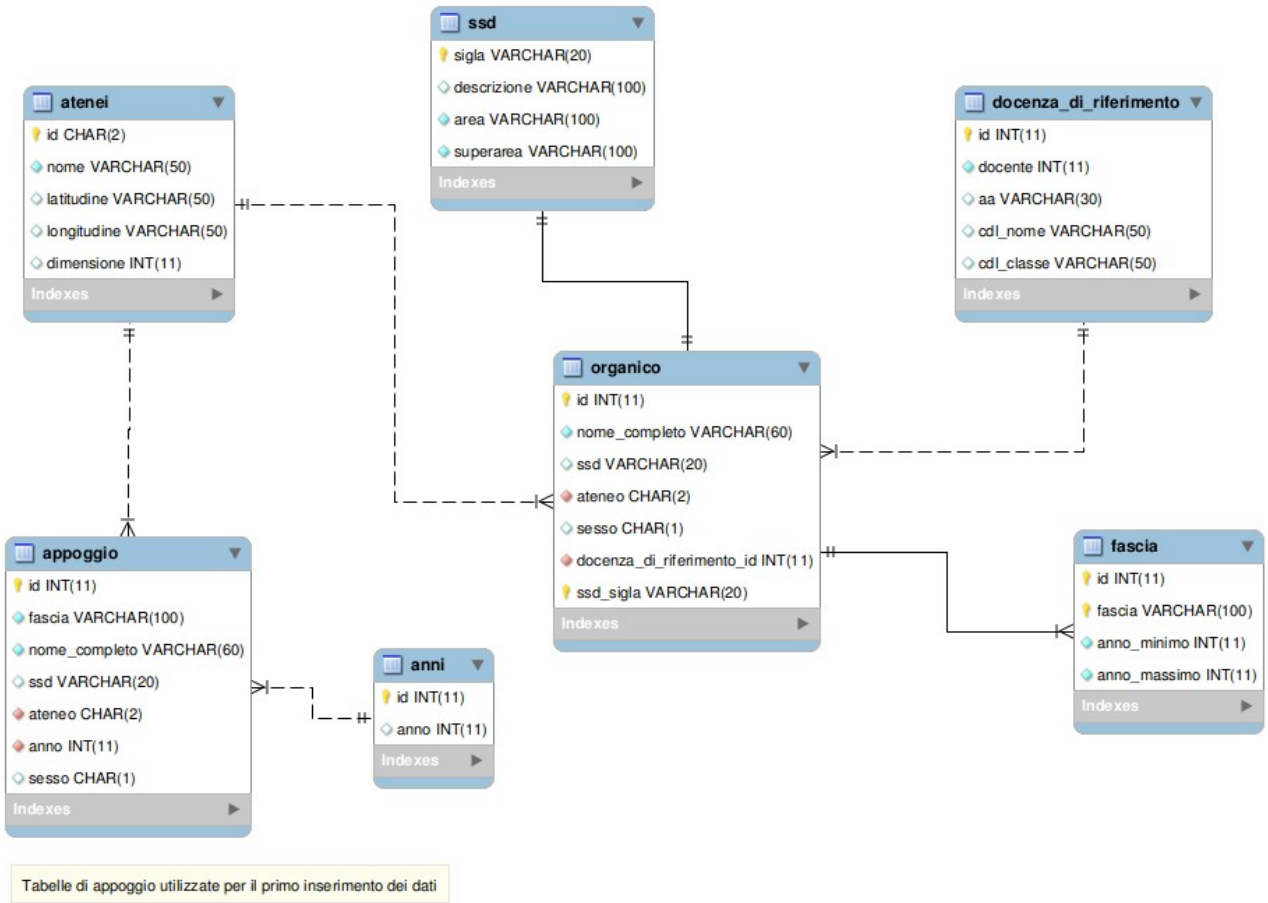
Le coordinate degli altri punti sono state espresse in termini delle coordinate dei primi tre, operando piccoli spostamenti per rendere la figura più realistica.

4. Il database

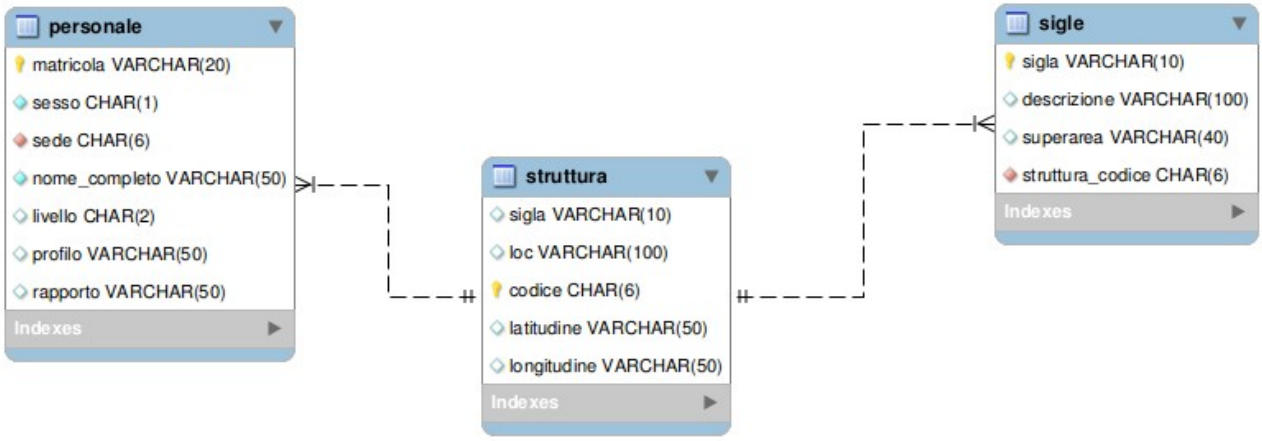
Il progetto si basa su due basi di dati che registrano la presenza femminile nell'organico delle università italiane e all'interno delle varie sedi del Consiglio Nazionale della Ricerca.

La struttura dei database è la seguente:

ORGANICO_MIUR



ORGANICO_CNR



4.1. I dati

Entrambi i database sono stati popolati per lo più automaticamente, scaricando i dati tramite curl dalle seguenti fonti:

1) http://cercauniversita.cineca.it/php5/docenti/cerca.php per ottenere i dati sull'anagrafica docenti e sulla fascia di abilitazione

2) http://www.universitaly.it/index.php/cercacorsi/universita per ottenere le schede SUA dei corsi di laurea attivi negli atenei italiani, in cui sono presenti i docenti designati come "docenti di riferimento" per un dato corso

3) http://statistica.miur.it/scripts/IU/vIU0_bis.asp per i dati relativi alla popolazione di ogni ateneo

4) http://www.miur.it/UserFiles/115.htm per la descrizione dei settori scientifico- disciplinari

5) http://www.dcp.cnr.it/DPUASI/ per i dati relativi al personale del CNR

Per integrare i dati sulla localizzazione degli atenei italiani e dei centri di ricerca sono state utilizzate le API di Google Maps. Le pagine risultanti dalle operazioni di estrazione sono state processate mediante le più comuni utility di bash (`tr`, `sed`, `awk`, `tail`, `grep`...). In seguito ai vari step

di lavoro, di cui parleremo tra qualche riga, abbiamo ottenuto i file in formato csv che sono stati automaticamente inseriti nel database.

4.1.1 MIUR

La creazione del database "`organico_miur`" ha presentato delle criticità rispetto alla referenziazione tra dati: questi, pur discendendo interamente dal sito del MIUR, provengono da fonti diverse. Per tale ragione sono emersi valori lievemente discordi e valori NULL in vari punti del database. Alcuni di questi sono stati risolti manualmente in seguito all'elaborazione automatica: ad esempio abbiamo inserito le geolocalizzazioni degli atenei lì dove le API di Google Maps hanno fallito. Per il resto, al fine di evitare il propagarsi di altri errori, abbiamo scelto di lasciare i valori così come risultavano nell'inserimento automatico.

4.1.1 a Anagrafica, SSD, Popolazione

Il sito non permette il download dell'intero database: quando si prova ad eseguire una ricerca troppo generica viene restituita la stringa "selezione troppo ampia". Abbiamo quindi scaricato i dati dalla pagina http://cercauniversita.cineca.it/php5/docenti/vis_docenti.php, alla quale, tramite POST, sono stati passati i parametri necessari per scaricare i dati relativi ad ogni ateneo e ad ogni anno.

(*rif. `scarica_organico.sh`)

I file risultanti presentavano un'intestazione da scartare sempre e contenevano all'interno di una tabella dati relativi all'organico. Abbiamo quindi ovviato al problema pulendo il codice html (*rif `step1.sh`) e aggiungendo il dato relativo all'ateneo e all'anno di riferimento.

A questo punto del lavoro, per ricavare il sesso del docente, è stato utilizzato l'applicativo Open Source `gender.c` (*rif. http://www.heise.de/ct/ftp/07/17/182/).
Questo applicativo permette di svolgere molte più operazioni di quelle utili ai nostri scopi: il fatto che il codice sorgente sia Open Source ci ha permesso di modificarlo lievemente eliminando funzioni superflue, rendendo possibile la lettura di più nomi in input, in modo da eseguire una sola chiamata per tutti i nomi presenti in ogni file, invece che di una chiamata per record.

L'applicativo, ad ogni istanza dell'input, assegna quindi un'etichetta tra:

- **F** > Feminine
- **M** > Masculine
- **U** > Unisex
- **E** > Error

Considerato che il formato dei nomi era della forma "`COGNOME Nome`", abbiamo passato all'applicativo la prima istanza come parametro, la cui forma era "`[A-Z][^A-Z]+`" (ovvero la prima stringa con iniziale maiuscola ma altre lettere minuscole), al fine di evitare errori sui doppi nomi: nei casi come "Paolo Maria", abbiamo pensato che considerare solo il primo nome avrebbe ridotto la percentuale di errori. (*rif. `step2.sh`). Inoltre, poiché la granularità della fasciazione ci sembrava sovrabbondante, abbiamo ridotto le fasce presenti alle seguenti classi: ordinario, associato, ricercatore, incaricato, straordinario, assistente. (*rif. `step3.sh`)

(*rif. `step1_miur_02_0` → `step2_miur_02_0` → `step3_miur_02_0` → `step4_miur_02_0`)

A questo punto, parallelamente agli scripts per la creazione del database e per l'inserimento dei dati circa i settori scientifico-disciplinari e gli atenei (*rif `organico_miur.sql`, `insert.sql`), è stato creato lo script sql per l'inserimento dei dati (*rif. `step4.sh` → `insert_step4.sql`). Tuttavia, nel momento in cui lo stesso docente è risultato in più file perché attivo nello stesso ateneo in diversi anni, sono emersi dei duplicati:

```
|623049 | Straordinario | ABATE Benedetto| GEO/02| 20| 2 | M
|625096 | Straordinario | ABATE Benedetto| GEO/02| 20| 3 | M
|629190 | Associato     | ABATE Benedetto| GEO/02| 20| 5 | M
|627143 | Straordinario | ABATE Benedetto| GEO/02| 20| 4 | M
|610767 | Ordinario     | ABATE Benedetto| GEO/02| 20| 10 | M
|612814 | Ordinario     | ABATE Benedetto| GEO/02| 20| 11 | M
|608720 | Ordinario     | ABATE Benedetto| GEO/02| 20|  1 | M
|637379 | Ordinario     | ABATE Benedetto| GEO/02| 20|  8 | M
|633284 | Ordinario     | ABATE Benedetto| GEO/02| 20|  7 | M
|639426 | Ordinario     | ABATE Benedetto| GEO/02| 20|  9 | M
```

Assumendo che un record con nome, ssd e ateneo uguali si riferisca allo stesso docente, abbiamo raggruppato i dati secondo questi ultimi parametri e li abbiamo redistribuiti dalla tabella “appoggio” alle tabelle “organico” e “fascia” (nota: se da una parte considerare l'ssd tra i parametri di raggruppamento ha ridotto gli errori sui casi di omonimia nello stesso ateneo – in questo modo Paolo Rossi fisico è diverso da Paolo Rossi storico – dall’altra ha introdotto duplicati nei dati iniziali in cui non era presente l'ssd per alcune delle annualità).

(*rif step5.sh → raggruppamento_appoggio.txt → elenco_organico.txt → insert_organico_fascia.sql)

Tramite le API di Google Maps è stata inserita l'informazione su latitudine e longitudine: con uno script bash (*rif atenei.sh → lista_atenei.txt → listacomandi.txt) è stata generata una stringa del formato di un array php e, tramite uno script php, è stato effettutato l'encoding dell'url da passare a curl (*rif google_geolocate.php → lista_comandi_php.txt). I risultati delle chiamate API, eseguite tramite uno script in bash (*rif esegui_comandi_php.sh), sono stati salvati in xml (*rif locate_01) e letti da un altro script che ha generato la lista dei comandi sql (*rif step6.sh → insert_latlng.sql).

Le informazioni relative alle superaree, ovvero ai settori scientifico-disciplinari (4), sono state mappate manualmente e i dati sono stati inseriti nel database sempre a partire da file csv. Per quanto riguarda i dati sulla popolosità degli atenei, è sopraggiunto il problema che nel file csv scaricato dalla fonte (3) non erano presenti gli stessi id usati come chiave primaria nella tabella “atenei”. Così è stato pulito il file accostando il più possibile le denominazioni degli atenei a quelle che plausibilmente erano già presenti nel database (*rif popolazione.csv → popolazione.sh → popolazione2.txt): l'inserimento è avvenuto tramite uno script php (*rif insert_atenei_pop.php) che risolve i casi di indecisione.

4.1.1.b Schede SUA

Dal portale (2) sono stati scaricati i dati relativi ai corsi di laurea universitari attivi negli atenei italiani, le cui schede SUA (Schede Uniche Annuali) registrano i nomi dei docenti che sono “docenti di riferimento” per un dato corso di laurea.

Le Url di riferimento sono tutte del tipo http://www.universitaly.it/index.php/scheda/sua/[0-9]{5} :

tuttavia molte di queste non contengono dati significativi. Dunque, non essendo riuscite a trovare una correlazione tra le cinque cifre dell'URL e le schede realmente esistenti, abbiamo scaricato tramite curl tutte le possibili combinazioni, eliminando successivamente (*rif `sua_step1.sh`) le pagine che non contenevano dati sulla docenza di riferimento (vuote o non esistenti). Dai file restanti abbiamo estrapolato l'informazione sul tipo di laurea descritto (nome, classe di laurea) ignorando l'ateneo per non generare cicli nel database e, come per il caso dell'anagrafica, abbiamo rimosso gli elementi html eccedenti. (*rif `sua_step2.sh` → `sua_step3.sh`)

(Esempio del processo di pulizia: *rif `step1_scheda_04872`, `step2_scheda_04872`, `step3_scheda_04872`.)

A questo punto si è reso necessario inserire i record nella tabella "`docenza_di_riferimento`", mappando il nome del docente da inserire con il suo id nella tabella organico (*rif `sua_step4.sh` → `load_sua.sql`). A causa dei duplicati generati dai passaggi di creazione della prima parte di database, è stato considerato come "id" il massimo risultato della query:

```
select max(id) from organico where nome_completo like concat
('%',@nome,' ',@cognome, '%')
```

E cioè l'id che presenta l'ssd lì dove esistono due o più record per lo stesso docente (nota: `@nome`, `@cognome` sono le colonne contenenti nome e cognome del docente nel file `step3_*`).

4.1.2 CNR

La realizzazione del database "`organico_cnr`" ha ricalcato la procedura già seguita per "`organico_miur`". Sono stati inseriti manualmente i dati relativi alla struttura e alla descrizione delle sigle degli istituti (*rif `insert.sql`). Tramite curl, sede per sede, i dati sono stati scaricati automaticamente dalla piattaforma (*rif `step1_cnr_220000`). A questo punto, esaminando solo i file che effettivamente contenevano dei dati, abbiamo pensato di considerare il personale per "sede di lavoro" piuttosto che per "sede di afferenza" e abbiamo così estratto i nominativi in questione (*rif `step1.sh` → `step2_cnr_220000`). A seguire, sono state eliminate le intestazioni delle tabelle e altri elementi html superflui, e il file è stato tramutato in formato csv (*rif `step2.sh` → `step3_cnr_220000`). Rispetto ai dati del MIUR, l'irregolare formattazione

di questi, ha comportato qualche complicazione nella pulizia degli elementi html e nell'estrapolazione dei dati puri. Pertanto, similmente a quanto fatto per l'organico universitario, è stato aggiunto il sesso ad ogni record, questa volta passando all'applicativo l'ultima stringa del nome (quindi “Maria” nel caso di “Paolo Maria”): in questo caso non c'era una differenziazione stilistica tra il cognome e il nome ed era difficile capire quale fosse, di volta in volta, l'ultima istanza di cognome e la prima di nome.

Contestualmente è stato generato lo script sql per l'inserimento dei dati nel database:

(*rif `step3.sh` → `step4_cnr_220000` → `insert_step4.sql`)

Qui, analogamente a quanto descritto nella sezione del MIUR, sono state effettuate le chiamate alle API di Google Maps e i valori risultanti sono stati aggiunti al database.

(*rif `google_geolocate.php` → `lista_comandi_curl` → `esegui_comandi_curl.sh` → `locate_ANCONA` → `step4.sh` → `instert_latlng.sql`)

4.2 – Limiti e prospettive del database

Nel processo di creazione del database sono state operate delle scelte che hanno portato all'esclusione di una porzione importante di informazione. Tra i “tagli” più importanti possiamo annoverare la divisione degli atenei tra università statali e libere, la suddivisione della popolazione di ogni ateneo rispetto al sesso, la regione di residenza di un dato ateneo (dato rilevante ai fini del diritto allo studio), ecc..

Ciononostante, restano memorizzati nel database molti dati non accessibili dall'interfaccia, che ne permette l'esplorazione fino ad un certo livello di granularità, ma non oltre. Non si può ad esempio personalizzare la ricerca sul settore scientifico-disciplinare, come non è possibile tra i corsi di laurea il cui nome contiene una data parola o tra i corsi di laurea che appartengono a una data classe di laurea. Probabilmente molti di questi dati non risultano rilevanti ma, non esistendo ad oggi statistiche complete in merito, non è possibile stabilire a priori quali correlazioni possano comportare modifiche percentuali sulla popolazione femminile. In prospettiva, sarebbe interessante sviluppare ulteriormente il database recuperando dati riguardanti altri livelli di istruzione e incrociandoli ad esempio con i dati Anvur.

5. Tecnologia Utilizzata

Sono stati adoperati i seguenti linguaggi:

- **HTML5** e **CSS3** per la realizzazione grafica dell'interfaccia
- **Javascript** per la gestione degli eventi lato client: nello specifico sono state utilizzate le librerie
- **jQuery**
- **d3** per la creazione di grafici svg per la visualizzazione
- **PHP5** per l'accesso al database
- **Apache2** come web server
- **MySQL** per la gestione del database
- **bash** per il pre-processing dei dati e per l'automatizzazione dell'attività di inserimento dei dati nel database e quindi di aggiornamento futuro dello stesso.