

**Fun facts**

500 million users (over 500 million according to Facebook's statistics, as of July 2010)

Linkedin: 45 million visits per day (based on Alexa estimates, as of 19/10/2010)

Se MySpace fosse una nazione sarebbe l'undicesima, tra Giappone e Brasile.

2 mesi di Youtube = 72 anni di contenuti di ABC, NBC e CBS insieme

**Data integration and data quality**

La Data Integration è definita in senso lato come la combinazione di dati provenienti da fonti diverse rispetto allo stesso o individuo o gruppo.



L'attuazione del record linkage:

75 per cento dello sforzo è nella preparazione del file di input

5 per cento dello sforzo è nella realizzazione del collegamento stesso

20 per cento dello sforzo è nel controllo dei risultati del collegamento



Raccolta delle informazioni sulle sorgenti dei dati

Investigazione preliminare

Target population and units

Capire i dati sorgente: i metadata

Implicazioni dai metadata



Preparare i dati per il record linkage

Errori tipici nelle linking variables

Standardisation: editing, parsing, formatting, concordance

Deduplication

Rendere anonimi i dati



### Editing

è il processo di rilevazione e il trattamento dei dati errati o sospetti

### Parsing and standardisation of linking variables

L'analisi di un campo separa le entità all'interno di quel campo per rendere il confronto più facile. Ad esempio, un nome di campo contenente il nome e il cognome sarebbero separati

### Standardisation of surnames and first names

Gli usi di base sono: primo, la sostituzione di ortografia e varianti di abbreviazione dei nomi e degli indirizzi che si verificano comunemente e, in secondo luogo, per usare le parole chiave generate durante il processo di standardizzazione come spunti per lo sviluppo delle subroutine di *parsing*.

### Phonetic coding

La codificazione fonetica è un modo di scrivere una stringa di caratteri in base al modo in cui la stringa è pronunciata, ed è uno strumento utile per riassumere i nomi e consentire alcune variazioni ortografiche



### Deduplication

L'analista (blacksmith) è in grado di eliminare i duplicati con un processo chiamato di deduplicazione. Può essere pensata anche come un esercizio di integrazione dei dati, in cui possono essere eseguite operazioni utilizzando le stesse tecniche di integrazione tra due file.

### Anonymisation of unique identifiers

Tenere presente le normative privacy.



### Kettle - Pentaho Data Integration

<http://pdi.pentaho.com>

ETL extract, transform, and load. As in extract data from a data source, transform it into what you need, and load it somewhere else.

### Talend Open Studio

download.Talend.com/Open

-Studio

### DQ Guru

[www.sqlpower.ca/page/dqguru](http://www.sqlpower.ca/page/dqguru)

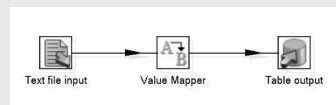
### Data cleaner

[datacleaner.eobjects.org/](http://datacleaner.eobjects.org/)



- Visual designer per la creazione delle trasformazioni
- Connettività per una miriade di banche dati, inclusi tutti i DB, formati di testo, ecc
- Supporta la distribuzione di jobs tra più server se si devono fare lavori seri
- Ottima gestione degli errori e sistemi di notifica di errore
- Comunità attiva (quella di Kettle un poco più ampia)
- Software gratuito e open source

- Steps: record stream
- Hops: connette gli steps



### Exact matching

VS the human approach

VS the statistical approach

Tramite algoritmi di matching è possibile combinare e comporre dati dissimili da sistemi sorgente multipli usando tecniche euristiche e vicinanze “fonetiche”.



### The m probability

Si tratta di una misura di quanto i dati sono attendibili, e può essere espressa come la 'probabilità che i due valori siano coincidenti, dato il fatto che si riferiscano alla stessa unità (ad esempio persona / azienda / ente / evento). Vale a dire:

$m = \Pr(\text{i due valori concordano} \mid \text{i record sono in matching})$

### The u probability

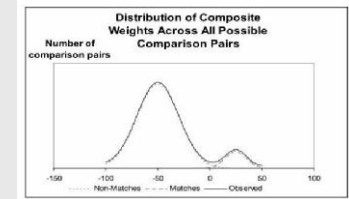
Il carattere comune del valore è descritto dalla 'probabilità u' o 'u prob'. Si tratta di una misura di quanto è probabile che i due valori saranno d'accordo per caso. Si esprime come 'la probabilità dei valori concordare dato che i record non si riferiscono alla stessa unità. Vale a dire:

$u = \Pr(\text{i due valori concordano} \mid \text{i record non sono un matching})$



### Weights

- ▣ Distribution
- ▣ Cut-off thresholds
- ▣ Clerical review



### Blocking

Per ridurre il numero di confronti effettuati e mettere a fuoco i record che hanno più probabilità di essere matching, i record possono essere filtrati prima in modo che solo determinati record sono considerati in rapporto gli uni agli altri.

### Passes

Un 'pass' è un'iterazione di record linkage che utilizza una combinazione di variabili blocking variables e variabili matching.



### Matching method

### Choice of blocking variables

### Choice of linking variables

- ▣ Commonly used comparison functions for linking variables

### Quality assessment of linked data

- ▣ False positives, false negatives and match rates

### Adding data over time



### Tools

Registry Plus™ Link Plus

[ftp://ftp.cdc.gov/pub/Software/RegistryPlus/Link\\_Plus/RLinkPlus-2.0.exe](ftp://ftp.cdc.gov/pub/Software/RegistryPlus/Link_Plus/RLinkPlus-2.0.exe)

FEBRL: <http://sourceforge.net/projects/febrl/>

### Chocemaker Analyzer

<http://oscm.sourceforge.net/index.html>

### Kettle

### Ddupe2 - Geoddupe

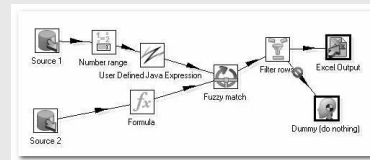
[www.cs.umd.edu/.../ddupe](http://www.cs.umd.edu/.../ddupe)

### FRIL

[fril.sourceforge.net](http://fril.sourceforge.net)

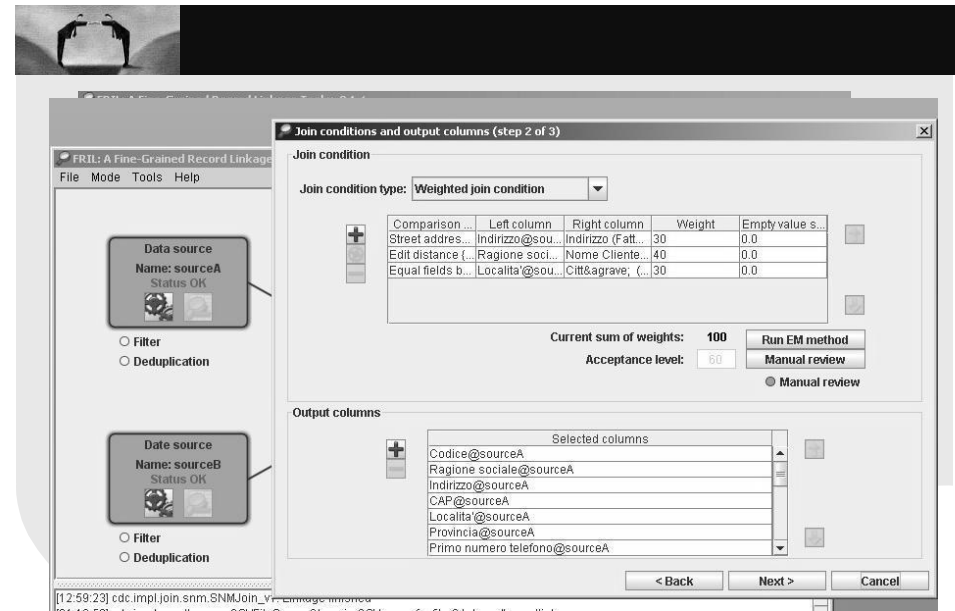
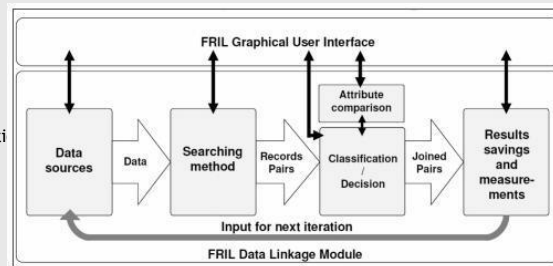


## Kettle - Pentaho Data Integration



## + FRIL

Fine-grained Record Integrati  
and Linkage tool



[12:59:23] cdc.impl.join.snm.SNMJoin\_V1 - Linkage finished  
[10:01:11] cdc.impl.resultsavers.CSVFileSaver: Close in CSV saver for file results.csv



## Business is ... Business

Il profiling e l'integrazione dei dati consente alle organizzazioni di identificare inesattezze e incongruenze di dati e per una migliore pianificazione ed esecuzione dei progetti.

Riduce i costi e migliora il ROI di attività di business strategiche quali il CRM, Data Warehouse e Business Intelligence, grazie a informazioni affidabili e valide.

