

IntegrAVIS

Sistema per l'integrazione
dei dati del donatore da ASSO-AVIS su ASSO-WEB

Francesco Trotta – DIEI Università Perugia





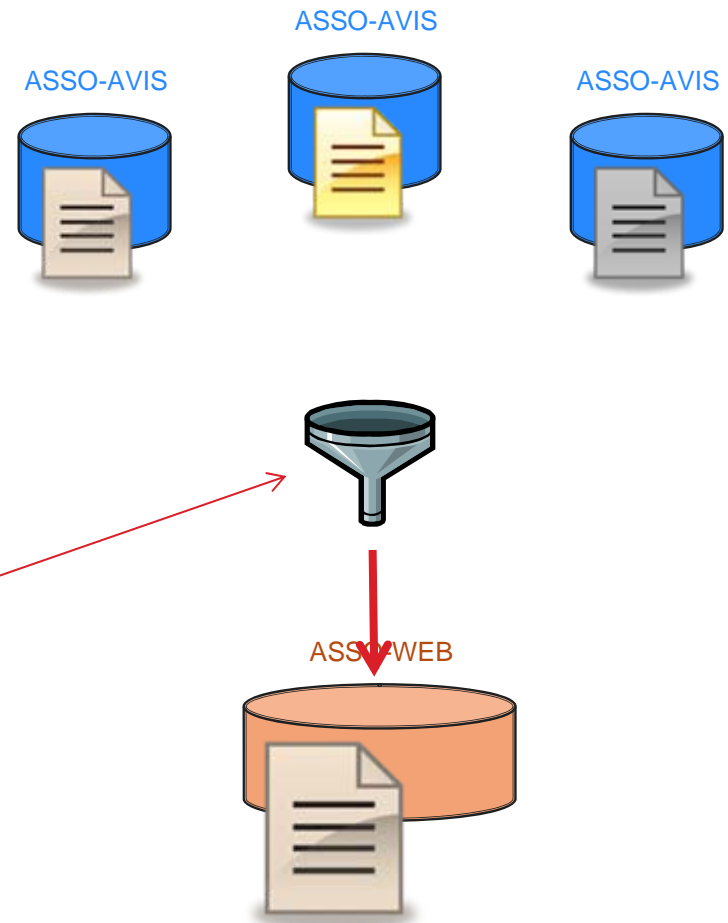
GIVI

Gruppo di Informatica e
Visualizzazione dell'Informazione

- Diretto dal prof. Giuseppe Liotta, 8 membri
- Presso il Dip. di Ingegneria Elettronica e dell'Informazione (DIEI) dell'Università di Perugia (S. Lucia)
- Competenze:
 - Progetto e sviluppo di sistemi per
 - Gestione e analisi di grandi quantità di dati (data mining)
 - Visualizzazione di dati relazionali (relational data analysis)
 - Visual Meta-search Clustering Web Engines (WhatsOnWeb)
 - Sicurezza informatica (sistemi per la fraud detection)
 - Human-Computer Interaction (interfacce diagrammatiche)
- Progetto *IntegrAVIS*:
 - Dr. Ing. Francesco Trotta (responsabile – francesco.trotta@diei.unipg.it)
 - Simone Liberali

Lo scenario della gestione dati donatore AVIS Umbria

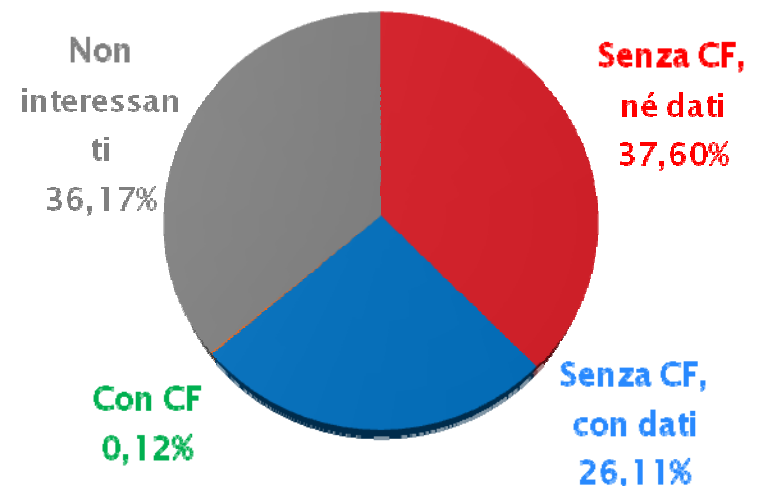
- Prima: ASSO-AVIS per la gestione dei dati relativi ai donatori, distribuito:
 - Ogni AVIS Comunale con propria copia di ASSO-Avis
 - Ogni AVIS Comunale gestisce autonomamente i propri dati
 - Adesso: ASSO-WEB (Tesi), centralizzato
 - Necessaria integrazione
 - Problemi:
 1. Dati parziali
 2. Dati non uniformi
 3. Dati duplicati
 4. Mappatura dati
- ASSO-AVIS ↔ ASSO-WEB?



Il profilo dei dati sorgente

- Più di 100 campi per donatore, con mappatura non campi sempre definita
- Totale record: 25.000 circa
- Perugia
 - 5872 totali
 - 3748 record interessanti
 - Identificazione del donatore:
 - Codice fiscale solo per 7 record
 - Dati per calcolo codice fiscale solo per 1533 record
- Dati parziali
- Dati non sempre uniformati
 - Refusi in nomi, cognomi, città, ...
 - Modifica delle province italiane
 - Politiche differenti per registrazione dati stranieri
- Medesimo donatore registrato in più Comunali

Profilo per l'identificazione del donatore (dati Perugia)



Obiettivi

1. Uniformare i dati presenti in ogni comunale
2. Definire un identificatore surrogato
3. Gestire i record duplicati
4. Definire un processo incrementale (dati non disponibili tutti insieme)
5. Realizzare base di dati unica uniformata ed integrata
6. Definire mappatura tra base di dati unica e ASSO-WEB

Analisi dello stato dell'arte

- Processi ETL
 - Extract: estrazione dei dati dalla sorgente
 - Transform: trasformazione dei dati estratti (uniformazione)
 - Load: caricamento dei dati nella base di dati di destinazione
- Processi ETL fanno parte dei sistemi di
 - Enterprise Resource Planning (ERP)
 - Customer Relationship Management (CRM)
- Strumenti esistenti
 - Censiti e verificati 14 strumenti, sia proprietari che OpenSource
- Il problema considerato:
 - Necessita di una misura della somiglianza basata su regole specifiche (no identificatore univoco)
 - Operazione *una tantum*

→ sistema *ad hoc*: IntegrAVIS

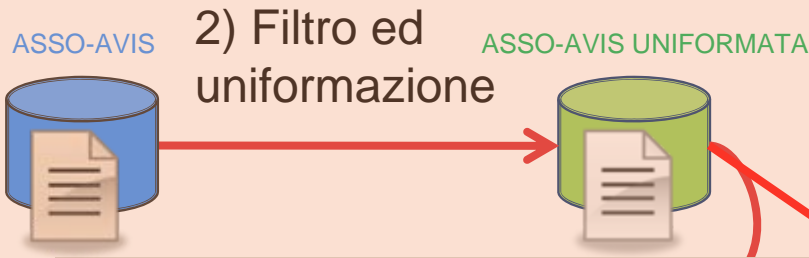
Le soluzioni adottate

1. Uniformazione dei dati
 - Algoritmo per gestione automatica dei refusi in città e province
 - L'algoritmo apprende con l'utilizzo
2. Rilevazione dei duplicati/mappatura
 - Identificatore surrogato:
 - Nome, cognome, data di nascita, sesso
 - Algoritmo per misurazione somiglianza identificatori (possibilità di refusi):
 - Distanza di Levenshtein adattiva
3. Dettagliato rapporto delle operazioni

Il processo complessivo

1) Prelievo base di dati (db) da Comunale

Ripetuto per ogni Comunale



4) Duplicati inter-db ed integrazione

ASSO-AVIS COMPLESSIVA

AVIS - Ricerca Duplicati

File Strumenti

Verifica Duplicati

Interrompi

Percentuale tuple completata: 17.61%

Provincia: CH

Cognome	Nome	Data di nascita	Sesso	Tessera	Città di nascita	Prov. Nascita	Città domicilio	Indirizzo Domicilio
	ANDREA	07/12/1969	M	292007895	NULL	PG	PERUGIA	VIA F. MAGELLANO 50
	ANDREA	07/12/1966	M	292009638	NULL	PG	PERUGIA	VIA DELLE GHIANDE 73
	LUISHIPOLITO	05/12/1972	M	292010033	NULL	PG	PERUGIA	VIA DEL GIOCHETTO (SEZ. DI MICROBIOLOGIA)
	LUISHIPOLITO	05/12/1972	M	292008096	PORTOGALLO	PG	PERUGIA	STRADA COLLESTRADA 24
	MARCO	15/01/1988	M	292020445	NULL	PG	PERUGIA	VIA DELLA FILARMONICA 27
	MARCO	17/01/1988	M	292009743	NULL	PG	PERUGIA	VIA FRA BEVIGNATE 34
	SILVANA	10/02/1958	F	292004927	MAGIONE	PG	PERUGIA	VIA BARONERICASOLI 25
	SILVANA	10/02/1956	F	292004562	PERUGIA	PG	CORCIANO	VIA .SETTEMBRINI 20/A/I

Mappatura ed integrazione

ASSO-WEB

Conclusioni

- Progettato ed implementato un sistema per l'integrazione dei dati del donatore da ASSO-AVIS ad ASSO-WEB
- Applicazione *ad hoc*
 - flessibilità di progettazione
 - no aggravio acquisto ulteriore software
- Processo automatico:
 - Uniformazione dati adattiva
 - Algoritmo *ad hoc* di stima della somiglianza per rilevazione dei duplicati
 - drastica riduzione dell'intervento umano
- Dettagliato rapporto delle operazioni
 - controllo completo da parte dell'operatore
- Processo incrementale:
 - Comunali possono essere gestite in sequenza (problemi logistico-organizzativi)
- Base algoritmica di molteplice applicazione