Master Universitario in: "ANALISI DATI PER LA BUSINESS INTELLIGENCE"

A.A. 2012 – 2013

Titolo della tesi: "Conservazione, riuso e disseminazione dei dati di inchiesta – Metodi e strumenti a confronto"

Autore: Paolo Dacorsi

Abstract

Scopo della presente dissertazione consiste nel mettere in luce la rilevanza di un sistema ben organizzato per l'archiviazione dei dati di inchiesta, che permetta di preservare il valore di tali dati e che ne faciliti il riuso e la disseminazione.

Un'archiviazione dei dati strutturata e funzionale presuppone un non indifferente lavoro in *back-end*, che prende avvio con la fase di progettazione della base dati (*database management*), la quale deve tener conto di alcuni accorgimenti metodologici:

-effettuare un'analisi dei dataset sorgente da inserire nell'archivio, che nell'ambito della ricerca sociale risultano tipicamente costituiti da dati grezzi e disaggregati (ovvero organizzati a livello del singolo individuo intervistato);

-valutare preventivamente l'obiettivo per cui la base dati viene creata, che consiste caratteristicamente nella predisposizione di un catalogo di dati e metadati liberamente (con l'ovvia esclusione dei dati sensibili relativi ai soggetti intervistati) ed agevolmente accessibile all'utente-ricercatore via web;

-considerare che per raggiungere l'obiettivo di una catalogazione di dati e metadati coerente e standardizzata è necessario che lo sia anche il processo di archiviazione stesso, che non può prescindere dalla definizione di uno standard per l'organizzazione in cartelle, per il flusso dei dati da un software all'altro (fino al deposito delle informazioni nella base dati) e per la prassi da seguire nei processi di archiviazione, un aspetto che risulta ancor più nevralgico nel caso in cui le operazioni di archiviazione vengano affidate a più *database manager* (com'è capitato durante la mia esperienza di stage, in quanto

l'organizzazione di un archivio dati survey ha assunto i connotati di un vero e proprio lavoro di *equipe*).

Alla progettazione della base dati fanno seguito, sempre dal lato *back-end*, operazioni di pulizia dei dati (*data quality*) - quali ad esempio l'assegnazione di formati alle variabili e la definizione di etichette -, finalizzate ad un agevole accesso ai dati via web (lato *front-end*), che consenta al ricercatore di raggiungere le informazioni di interesse con semplicità e nel minor tempo possibile, esigenze non troppo dissimili da quelle che in un contesto aziendale sarebbero alla base dell'approccio della *business intelligence*. Nel caso specifico dei dati di inchiesta, è bene tener presente che:

-il numero di variabili caricate nell'archivio dati può risultare particolarmente ingente, di conseguenza il ricercatore sfrutterà spesso e volentieri funzioni di ricerca e ricerca avanzata per item, che per funzionare correttamente devono essere supportate da un'attenta definizione dei nomi e delle etichette delle variabili, che assicuri di individuare le informazioni di interesse senza che venga ingenerata confusione;

-la medesima domanda ripetuta in diverse wave successive di un'indagine deve essere facilmente riconducibile alle sue "gemelle", in modo tale da permettere al ricercatore di apprezzare l'evoluzione nel tempo del fenomeno di interesse.

Dal lato *front-end*, su cui si focalizzerà in modo particolare la presente tesi, è auspicabile che l'interfaccia web permetta non soltanto di ricercare i dati di interesse in modo funzionale ed "a distanza" (vale a dire sganciata dall'onere di contattare il disseminatore dei dati ogni qual volta si sia interessati a svolgere analisi sui dati), ma anche di condurre sulle variabili di interesse elaborazioni statistiche (*analytics*), quali ad esempio distribuzioni di frequenza bivariate e semplici rappresentazioni grafiche (istogrammi di frequenza, grafici a torta e diagrammi a dispersione), pur tenendo conto del fatto che l'accesso alla base dati via web può non permettere di "incrociare" dati provenienti da diversi dataset, come nel caso del software Nesstar.

Se la prima parte della trattazione avrà le fattezze di una carrellata dei più illustri esempi di catalogazione dei dati di inchiesta, nel prosieguo verrà messa in luce l'evoluzione dei sistemi di archiviazione, presentando nel dettaglio le funzionalità dei software BIAS e Nesstar.

Verrà infine discusso l'utilizzo del software SAS per organizzare un archivio di dati survey.