Come archiviare i dati per le scienze sociali

ADPSS-SOCIODATA

Archivio Dati e Programmi per le Scienze Sociali
www.sociologiadip.unimib.it/sociodata
E-mail: adpss.sociologia@unimib.it
Tel.: 02 64487513
Fax: 02 64487561

Perché ADPSS-Sociodata richiede dati ben documentati?

Una buona documentazione dei dati archiviati rappresenta un vantaggio sia dal punto di vista della loro reperibilità, sia da quello della loro utilizzabilità per la ricerca sociale. La possibilità di realizzare analisi secondaria di dati per le scienze sociali è, infatti, legata alla disponibilità di archivi di dati e meta-dati predisposti secondo criteri concordati a livello internazionale. La collaborazione dei ricercatori alla costruzione di tali archivi è condizione indispensabile alla crescita quantitativa e qualitativa dell'attività di ricerca scientifica.

Cosa dovrebbe essere fornito a ADPSS-Sociodata?

Ci sono tre principali tipi di materiali che costituiscono la documentazione ideale per un *data-set*:

- **Materiale esplicativo**: è indispensabile per un uso scientificamente corretto di un *data-set* in quanto, senza di esso, non può essere raggiunta una piena comprensione dei dati. L'assenza del materiale esplicativo compromette la possibilità di archiviare un *data-set*.

- **Informazioni di contesto**: servono agli utenti per conoscere il contesto in cui i dati sono stati raccolti e l'utilizzo a cui essi erano originariamente destinati. Le informazioni di contesto permettono inoltre di costruire una memoria storica dei dati più chiara e dettagliata. Sebbene non essenziale, la presenza di queste informazioni è fortemente raccomandata.

- **Informazione circa la catalogazione dei dati**: la possibilità di creare una documentazione standard di descrizione delle indagini è ovviamente legata alla disponibilità di questo materiale. Le informazioni sulla catalogazione hanno due finalità principali: in primo luogo, servono come documentazione dei *data-set* utile alla loro citazione bibliografica nell'ambito di pubblicazioni che derivano da analisi secondarie di tali dati; inoltre, rappresentano la principale fonte informativa utilizzata per la costruzione degli archivi di meta-dati. Questo materiale è raccolto attraverso la *scheda di descrizione delle indagini* predisposta da ADPSS-Sociodata.

Materiale esplicativo

Questo materiale dovrebbe raccogliere le informazioni basilari che, in linea teorica, andrebbero fornite insieme al *data-set* da chi ha realizzato l'indagine. Gli aspetti che devono essere documentati sono:

1. Il metodo di raccolta dei dati. Le procedure utilizzate per la raccolta di dati vanno dettagliatamente descritte. Nel caso di indagini campionarie, sono di primaria importanza le informazioni sul disegno campionario e sulle procedure utilizzate per l'estrazione del campione. Risulta poi di grande utilità documentare la presenza di ricerche pilota, di procedure di monitoraggio realizzate nel corso della raccolta di dati e, inoltre, di ogni altro controllo di qualità dei dati realizzato. Anche la copertura temporale e spaziale del *data-set* deve essere specificata.

2. La struttura del data-set. Vanno indicati il numero di casi e delle variabili presenti in ciascuno dei *file* e il numero di *files* che compongono il *data-set*. Nel caso in cui il *data-set* si componga di più basi di dati logicamente interrelate (ad esempio se sono presenti due basi di dati di cui la prima comprende informazioni rilevate sulle famiglie mentre la seconda informazioni rilevate su individui appartenenti alle stesse famiglie), è indispensabile produrre un diagramma che indichi l'insieme di relazioni intercorrenti tra le basi di dati.

3. Il supporto tecnico. Sono le informazioni relative al software e al sistema operativo utilizzati per creare i *files*. Dovrebbero anche comprendere una lista completa con i nomi dei *files* che compongono il *data-set*.

4. Le variabili, gli schemi di codifica e di classificazione. Si tratta, in primo luogo, dalla lista completa delle variabili presenti nel data-set, con la specificazione di tutte le codifiche e classificazioni utilizzate. E' inoltre utile evidenziare l'eventuale presenza di variabili che si rifanno a schemi standard di codifica e di classificazione.

5. Gli indicatori. La costruzione di indicatori a partire dai dati originariamente raccolti è un'operazione molto comune, che va documentata in modo dettagliato. Nei casi più semplici - ad esempio per età raggruppare all'interno di classi - può essere sufficiente spiegare la logica di costruzione dell'indicatore attraverso le etichette applicate alle modalità che lo compongono ("meno di 20 anni"; "da 21 a 40 anni"; …). Nei casi più complessi, sono necessarie altre modalità di documentazione, in grado di chiarire lo schema logico del procedimento che ha generato gli indicatori: il modo migliore è quello di fornire, insieme al data-set, i programmi che hanno generato gli indicatori, come ad esempio i file di sintassi di *SPSS* o di *STATA*. Un esempio

di sintassi SPSS che esplica i criteri con i quali è stato costruito un indicatore del livello di consumo mediatico a partire da due variabili dicotomiche può essere:

```
if (VAR1=0)&(VAR2=0) VAR3=0.
if (VAR1=0)&(VAR2=1) VAR3=1.
if (VAR1=1)&(VAR2=0) VAR3=1.
if (VAR1=1)&(VAR2=1) VAR3=2.
var lab VAR3 "livello di consumo mediatico".
val lab VAR3
0 "basso"
1 "medio"
2 "alto".
EXE.
```

6. Le variabili di ponderazione (i "pesi"). Queste variabili devono essere documentate in modo completo, indicandone le modalità di costruzione e le circostanze in cui vanno utilizzate. Questo ultimo aspetto è necessario soprattutto se sono presenti nel *data-set* una moltitudine di pesi, ognuno dei quali ha uno scopo differente. Ad esempio, per un'indagine trans-nazionale in cui sono disponibili due pesi – quello di "disegno" per correggere la distorsione connessa al disegno campionario adottato e quello di "popolazione" per correggere la distorsione che deriva dalla differente numerosità delle popolazioni nazionali rilevate - può essere fornita una tabella del seguente tipo:

Obiettivo conoscitivo	***Esempio*: partecipazione al voto (% dei rispondenti che hanno votato alle ultime elezioni)**	***Pesi da usare:***	
		Peso di "disegno"	**Peso di "popolazione"**
a) Analizzare i dati di un solo paese	i) Partecipazione al voto in Italia	X	
	ii) Partecipazione al voto in Italia per età e genere	X	
b) Comparare i risultati di due o più paesi, senza fare riferimento a valori totali o medi	Comparare la partecipazione al voto in Italia e Francia	X	
c) Analizzare i dati di due o più paesi considerati congiuntamente	i) Partecipazione al voto in Scandinavia	X	X
	ii) Partecipazione al voto nell'UE	X	X
	iii) Partecipazione al voto in UE per età e sesso	X	X

7. Il questionario. Quando i dati provengono dalle risposte date ad un questionario, il testo di ciascuna domanda dovrebbe essere riportato nella stessa versione utilizzata per la rilevazione dei dati. L'ideale è che al testo di ogni domanda sia associato un riferimento alla variabile che ha generato. Ad esempio, una domanda potrebbe essere così riportata nel questionario:

V12 in genere, per quanto tempo al giorno guarda telegiornali o programmi di politica e attualità alla televisione?

Mai	00
Meno di ½ ora	01
½ ora -1 ora	02
Più di 1 ora - 1 ora e ½	03
Più di 1ora e ½ - 2 ore	04
Più di 2 ore - 2 ore e ½	05
Più di 2 ore e ½- 3 ore	06
Più di 3 ore	07
(non so)	88

---

8. La privacy. E' importante specificare se i dati contengono informazioni confidenziali relative a individui, famiglie, organizzazioni o istituzioni. Tali informazioni dovrebbero essere rimosse o rese anonime prima di fornire il *data-set* ad ADPSS-Sociodata.

9. La convalida dei dati e altri controlli. Sono queste le informazioni su eventuali errori nei dati e sulle procedure di controllo o "ripulitura" realizzate nel corso della loro rilevazione o successivamente.

Informazioni di contesto

Questa documentazione completa il materiale esplicativo di base aggiungendo ricchezza e profondità alle informazioni già disponibili. In particolare, la documentazione riguarda i seguenti aspetti:

1. La descrizione del progetto. Si tratta di informazioni sull'origine e sullo sviluppo del progetto e del *data-set* a livello intellettuale, finanziario e organizzativo. Per esempio, è necessario specificare i motivi alla base della raccolta dei dati, gli scopi del progetto, le pubblicazioni previste e ogni altra rilevante informazione di contesto.

2. La provenienza del data-set. Si tratta di tutte quelle informazioni utili ad la ricostruzione dettagliata delle procedure di raccolta dei dati. Tali informazioni sono di particolare rilievo nel caso in cui siano intervenuti cambiamenti nei metodi utilizzati.

3. I data-set longitudinali. Per le indagini *cross-sectional* (ripetute nel tempo su campioni diversi), le indagini *panel* (ripetute nel tempo sugli stessi campioni) o per i dati aggregati raccolti all'interno di serie storiche, è particolarmente importante avere a disposizione informazioni aggiuntive sulle variazioni organizzative e metodologiche avvenute nel tempo: ad esempio, sui cambiamenti nel contenuto della variabile, nel testo della domanda, nella classificazione della variabile o nelle procedure di campionamento.

Informazioni circa la catalogazione dei dati

La maggior parte delle informazioni necessarie all'archiviazione dei dati sono richieste all'interno della *scheda di descrizione delle indagini*. Nella scheda vengono richieste informazioni come il titolo del *data-set*, i responsabili dell'indagine e delle rilevazioni, i tempi e i metodi della raccolta dei dati, il disegno e lo schema campionario, la copertura temporale e geografica dei dati. Ulteriori informazioni utili alla catalogazione dei *data-set* possono essere estrapolate da pubblicazioni o articoli ad esso riferiti.